AB

PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 :
C12N 15/12, C07K 14/47, 14/705, C12Q 1/68, C07K 16/18, 16/28, A61K 38/17, G01N 33/68

A2

(11) International Publication Number: **WO 99/33981**

(43) International Publication Date: 8 July 1999 (08.07.99)

(21) International Application Number: PCT/US98/27598

(22) International Filing Date: 22 December 1998 (22.12.98)

(30) Priority Data:
09/002,485 31 December 1997 (31.12.97) US

(63) Related by Continuation (CON) or Continuation-in-Part (CIP) to Earlier Application
US 09/002,485 (CIP)
Filed on 31 December 1997 (31.12.97)

(71) Applicant *(for all designated States except US)*: INCYTE PHARMACEUTICALS, INC. [US/US]; 3174 Porter Drive, Palo Alto, CA 94304 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: LAL, Preeti [IN/US]; 2382 Lass Drive, Santa Clara, CA 95054 (US). HILLMAN, Jennifer, L. [US/US]; 230 Monroe Drive #12, Mountain View, CA 94040 (US). CORLEY, Neil, C. [US/US]; 1240 Dale Avenue #30, Mountain View, CA 94040 (US). GUEGLER, Karl, J. [CH/US]; 1048 Oakland Avenue, Menlo Park, CA 94025 (US). BAUGHN, Mariah, R. [US/US]; 1537 LaRossa Circle, San Jose, CA 95125 (US). SATHER, Susan, K. [US/US]; Apartment #203, 2721 Midtown Court, Palo Alto, CA 94303 (US). SHAH, Purvi [IN/US]; 1608 Queen Charlotte Drive #5, Sunnyvale, CA 94087 (US).

(74) Agents: BILLINGS, Lucy, J. et al.; Incyte Pharmaceuticals, Inc., 3174 Porter Drive, Palo Alto, CA 94304 (US).

(81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GE, GH, GM, HR, HU, ID, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published
Without international search report and to be republished upon receipt of that report.

(54) Title: HUMAN SIGNAL PEPTIDE–CONTAINING PROTEINS

(57) Abstract

The invention provides a human signal peptide–containing proteins (SIGP) and polynucleotides which identify and encode SIGP. The invention also provides expression vectors, host cells, antibodies, agonists, and antagonists. The invention also provides methods for treating or preventing disorders associated with expression of SIGP.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece			TR	Turkey
BG	Bulgaria	HU	Hungary	ML	Mali	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MN	Mongolia	UA	Ukraine
BR	Brazil	IL	Israel	MR	Mauritania	UG	Uganda
BY	Belarus	IS	Iceland	MW	Malawi	US	United States of America
CA	Canada	IT	Italy	MX	Mexico	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NE	Niger	VN	Viet Nam
CG	Congo	KE	Kenya	NL	Netherlands	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NO	Norway	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	NZ	New Zealand		
CM	Cameroon			PL	Poland		
CN	China	KR	Republic of Korea	PT	Portugal		
CU	Cuba	KZ	Kazakstan	RO	Romania		
CZ	Czech Republic	LC	Saint Lucia	RU	Russian Federation		
DE	Germany	LI	Liechtenstein	SD	Sudan		
DK	Denmark	LK	Sri Lanka	SE	Sweden		
EE	Estonia	LR	Liberia	SG	Singapore		

HUMAN SIGNAL PEPTIDE-CONTAINING PROTEINS

TECHNICAL FIELD

This invention relates to nucleic acid and amino acid sequences of human signal peptide-containing proteins and to the use of these sequences in the diagnosis, treatment, and prevention of cancer and immunological disorders.

BACKGROUND OF THE INVENTION

Protein transport is an essential process for all living cells. Transport of an individual protein usually occurs via an amino-terminal signal sequence which directs, or targets, the protein from its ribosomal assembly site to a particular cellular or extracellular location. Transport may involve any combination of several of the following steps: contact with a chaperone, unfolding, interaction with a receptor and/or a pore complex, addition of energy, and refolding. Moreover, an extracellular protein may be produced as an inactive precursor. Once the precursor has been exported, removal of the signal sequence by a signal peptidase and posttranslational processing (e.g., glycosylation or phosphorylation) activates the protein. Signal sequences are common to receptors, matrix molecules (e.g., adhesion, cadherin, extracellular matrix, integrin, and selectin), cytokines, hormones, growth and differentiation factors, neuropeptides, vasomediators, phosphokinases, phosphatases, phospholipases, phosphodiesterases, G and Ras-related proteins, ion channels, transporters/pumps, proteases, and transcription factors.

G-protein coupled receptors (GPCRs) are a superfamily of integral membrane proteins which transduce extracellular signals. GPCRs include receptors for biogenic amines, e.g., dopamine, epinephrine, histamine, glutamate (metabotropic effect), acetylcholine (muscarinic effect), and serotonin; for lipid mediators of inflammation such as prostaglandins, platelet activating factor, and leukotrienes; for peptide hormones such as calcitonin, C5a anaphylatoxin, follicle stimulating hormone, gonadotropin releasing hormone, neurokinin, oxytocin, and thrombin; and for sensory signal mediators, e.g., retinal photopigments and olfactory stimulatory molecules.

The structure of these highly-conserved receptors consists of seven hydrophobic

transmembrane regions, cysteine disulfide bridges between the second and third extracellular loops, an extracellular N-terminus, and a cytoplasmic C-terminus. Three extracellular loops alternate with three intracellular loops to link the seven transmembrane regions. The N-terminus interacts with ligands, the disulfide bridge interacts with agonists and antagonists, and the large third intracellular loop interacts with G proteins to activate second messengers such as cyclic AMP (cAMP), phospholipase C, inositol triphosphate, or ion channel proteins. The most conserved parts of these proteins are the transmembrane regions and the first two cytoplasmic loops. A conserved, acidic-Arg-aromatic triplet present in the second cytoplasmic loop may interact with the G proteins. The consensus pattern, [GSTALIVMYWC]-[GSTANCPDE]-{EDPKRH}-x(2)-[LIVMNQGA]-x(2)-[LIVMFT]-[GSTANC]-[LIVMFYWSTAC]-[DENH]- R-[FYWCSH]-x(2)-[LIVM] is characteristic of most proteins belonging to this superfamily. (Watson, S. and Arkinstall, S. (1994) The G-protein Linked Receptor Facts Book, Academic Press, San Diego, CA, pp. 2-6; and Bolander, F.F. (1994) Molecular Endocrinology, Academic Press, San Diego, CA, pp. 8-19.)

Tetraspanins are a superfamily of membrane proteins which facilitate the formation and stability of cell-surface signaling complexes containing lineage-specific proteins, integrins, and other tetraspanins. They are involved in cell activation, proliferation (including cancer), differentiation, adhesion, and motility. These proteins cross the membrane four times, have conserved intracellular N- and C-termini and an extracellular, non-conserved hydrophilic domain. Three highly conserved polar amino acids are located in the transmembrane domains (TM), an asparagine in TM1 and a glutamate or glutamine in TM3 and TM4. Two to three conserved charged residues, including a glutamic acid residue, are present in the cytoplasmic loop between TM2 and TM3. The extracellular loop between TM3 and TM4 contains four conserved cysteine residues: two in a conserved CCG motif located about 50 residues C-terminal to TM3; one, often preceded by glycine, 11 residues N-terminal to TM4; and one in the extracellular loop may be found in a PXSC motif. Tetraspanins include, e.g., platelet and endothelial cell membrane proteins, leukocyte surface proteins, tissue specific and tumorous antigens, and the retinitis pigmentosa-associated gene peripherin. (Maecker, H.T. et al. (1997) FASEB J. 11:428-442.) Matrix proteins (Mps) function in formation, growth, remodeling and maintenance

of tissues and as important mediators and regulators of the inflammatory response. The expression and balance of MPs may be perturbed by biochemical changes that result from congenital, epigenetic, or infectious diseases. In addition, MPs affect leukocyte migration, proliferation, differentiation, and activation in immune response.

MPs encompass a variety of proteins and their functions. Extracellular matrix (ECM) proteins are multidomain proteins that play an important role in the diverse functions of the ECM. ECM proteins are frequently characterized by the presence of one or more domains which may include collagen-like domains, EGF-like domains, immunoglobulin-like domains, fibronectin-like domains, vWFA-like modules. (Ayad, S. et al. (1994) The Extracellular Matrix Facts Book, Academic Press, San Diego, CA, pp. 2-16.) Cell adhesion molecules (CAMs) have been shown to stimulate axonal growth through homophilic and/or heterophilic interactions with other molecules. In addition, interactions between adhesion molecules and their receptors can potentiate the effects of growth factors upon cell biochemistry via shared signaling pathways. (Ruoslahti, E. (1997) Kidney Int. 51:1413-1417.) Cadherins comprise a family of calcium-dependant glycoproteins that function in mediating cell-cell adhesion in solid tissues of multicellular organisms. Integrins are ubiquitous transmembrane adhesion molecules that link cells to the ECM by interacting with the cytoskeleton. Integrins also function as signal transduction receptors and stimulate changes in intracellular calcium levels and protein kinase activity. (Sjaastad, M.D. and Nelson, W.J. (1997) BioEssays 19:47-55.) Lectins are proteins characterized by their ability to bind carbohydrates on cell membranes by means of discrete, modular carbohydrate recognition domains, CRDs. (Kishore, U. et al. (1997) Matrix Biol. 15:583-592.) Certain cytokines and membrane-spanning proteins have CRDs which may enhance interactions with extracellular or intracellular ligands, with proteins in secretory pathways, or with molecules in signal transduction pathways. The lipocalin superfamily constitutes a phylogenetically conserved group of more than forty proteins that function by binding to and transporting a variety of physiologically important ligands. Members of this family function as carriers of retinoids, odorants, chromophores, pheromones, and sterols, and a subset of these proteins may be multifunctional, serving as either a biosynthetic enzyme or as a specific enzyme inhibitor. (Tanaka, T. et al. (1997) J. Biol. Chem. 272:15789-15795; and van't Hof, W. et al. (1997) J. Biol. Chem. 272:1837-1841.) Selectins are a family of calcium ion-dependent lectins

expressed on inflamed vascular endothelium and the surface of some leukocytes. They mediate rolling movement and adhesive contacts between blood cells and blood vessel walls. The structure of the selectins and their ligands supports the type of bond formation and dissociation that allows a cell to roll under conditions of flow. (Rossiter, H. et al. (1997) Mol. Med. Today 3:214-222.)

Protein kinases regulate many different cell proliferation, differentiation, and signaling processes by adding phosphate groups to proteins. Reversible protein phosphorylation is a key strategy for controlling protein functional activity in eukaryotic cells. The high energy phosphate which drives this activation is generally transferred from adenosine triphosphate molecules (ATP) to a particular protein by protein kinases and removed from that protein by protein phosphatases. Phosphorylation occurs in response to extracellular signals, cell cycle checkpoints, and environmental or nutritional stresses. Protein kinases may be roughly divided into two groups; protein tyrosine kinases (PTKs) which phosphorylate tyrosine residues, and serine/threonine kinases (STKs) which phosphorylate serine or threonine residues. A few protein kinases have dual specificity. A majority of kinases contain a similar 250-300 amino acid catalytic domain which can be further divided into eleven subdomains. The N-terminal domain, which contains subdomains I to IV, generally folds into a two-lobed structure which binds and orients the ATP (or GTP) donor molecule. The larger C terminal domain, which contains subdomains VIA to XI, binds the protein substrate and carries out the transfer of the gamma phosphate from ATP to the hydroxyl group of the target amino acid residue. Subdomain V links the two domains. Each of the 11 subdomains contain specific residues and motifs that are characteristic and are highly conserved. (Hardie, G. and Hanks, S. (1995) The Protein Kinase Facts Book, Vol I, pp. 7-47, Academic Press, San Diego, CA.)

Protein phosphatases remove phosphate groups from molecules previously modified by protein kinases thus participating in cell signaling, proliferation, differentiation, contacts, and oncogenesis. Protein phosphorylation is a key strategy used to control protein functional activity in eukaryotic cells. The high energy phosphate is transferred from ATP to a protein by protein kinases and removed by protein phosphatases. There appear to be three, evolutionarily-distinct protein phosphatase gene families: protein phosphatases (PPs); protein tyrosine phosphatases (PTPs); and

acid/alkaline phosphatases (APs). PPs dephosphorylate phosphoserine/threonine residues and are an important regulator of many cAMP mediated, hormone responses in cells. PTPs reverse the effects of protein tyrosine kinases and therefore play a significant role in cell cycle and cell signaling processes. Although APs dephosphorylate substrates in vitro, their role in vivo is not well known. (Carbonneau, H. and Tonks, N.K. (1992) Annu. Rev. Cell Biol. 8:463-493.)

Protein phosphatase inhibitors control the activities of specific phosphatases. A specific inhibitor of PP-I, I-1, has been identified that when phosphorylated by cAMP-dependent protein kinase (PKA) specifically binds to PP-I and inhibits its activity. Since PP-I is dephosphoryles many of the proteins phosphorylated by PKA, activation of I-1 by PKA serves to amplify the effects of PKA and the many cAMP-dependent responses mediated by PKA. In addition, since PP-I also dephosphorylates many phosphoproteins that are not phosphorylated by PKA, I-1 activation serves to exert cAMP control over other protein phosphorylations. I_1PP2A is a specific and potent inhibitor of PP-IIA. (Li, M. et al. (1996) Biochemistry 35:6998-7002.) Since PP-IIA is the main phosphatase responsible for reversing the phosphorylations of serine/threonine kinases, I_1PP2A has broad effects in controlling protein phosphorylations.

Cyclic nucleotides (cAMP and cGMP) function as intracellular second messengers to transduce a variety of extracellular signals, including hormones, and light and neurotransmitters. Cyclic nucleotide phosphodiesterases (PDEs) degrade cyclic nucleotides to their corresponding monophosphates, thereby regulating the intracellular concentrations of cyclic nucleotides and their effects on signal transduction. At least seven families of mammalian PDEs have been identified based on substrate specificity and affinity, sensitivity to cofactors and sensitivity to inhibitory drugs. (Beavo, J.A. (1995) Physiological Reviews 75: 725-748.) PDEs are composed of a catalytic domain of ~270 amino acids, an N-terminal regulatory domain responsible for binding cofactors and, in some cases, a C-terminal domain with unknown function. Within the catalytic domain, there is approximately 30% amino acid identity between PDE families and ~85-95% identity between isozymes of the same family. Furthermore, within a family there is extensive similarity (>60%) outside the catalytic domain, while across families there is little or no sequence similarity. A variety of diseases have been attributed to increased PDE activity and inhibitors of PDEs have been used effectively as anti-inflammatory,

antihypertensive, and antithrombotic agents. (Verghese, M.W. et al. (1995) Mol. Pharmacol. 47:1164-1171; and Banner, K.H.. and Page, C.P. (1995) Eur. Respir. J. 8:996-1000.)

Phospholipases (PLs) are enzymes that catalyze the removal of fatty acid residues from phosphoglycerides. PLs play an important role in transmembrane signal transduction and are named according to the specific ester bond in phosphoglycerides that is hydrolyzed, i.e., A_1, A_2, C or D. PLA_2 cleaves the ester bond at position 2 of the glycerol moiety of membrane phospholipids giving rise to arachidonic acid. Arachidonic acid is the common precursor to four major classes of eicosanoids; prostaglandins, prostacyclins, thromboxanes and leukotrienes. Eicosanoids are signaling molecules involved in the contraction of smooth muscle, platelet aggregation, and pain and inflammatory responses. PLC is an important link in certain receptor-mediated, signaling transduction pathways. Extracellular signaling molecules including hormones, growth factors, neurotransmitters, and immunoglobulins bind to their respective cell surface receptors and activate PLC. Activated PLC generates second messenger molecules from the hydrolysis of inositol phospholipids that regulate cellular processes, e.g., secretion, neural activity, metabolism and proliferation. (Alberts, B. et al. (1994) Molecular Biology of The Cell, Garland Publishing, Inc., New York, NY, pp. 85, 211, 239-240, 642-645.)

The nucleotide cyclases, i.e., adenylate and guanylate cyclase, catalyze the synthesis of the cyclic nucleotides, cAMP and cGMP, from ATP and GTP, respectively. They act in concert with phosphodiesterases, which degrade cAMP and cGMP, to regulate the cellular levels of these molecules and their functions. cAMP and cGMP function as intracellular second messengers to transduce a variety of extracellular signals, e.g., hormones, and light and neurotransmitters. Adenylate cyclase is a plasma membrane protein that is coupled with various hormone receptors also located on the plasma membrane. Binding of a hormone to its receptor activates adenylate cyclase which, in turn, increases the levels of cAMP in the cytosol. The activation of other molecules by cAMP leads to the cellular effect of the hormone. In a similar manner, guanylate cyclase participates in the process of visual excitation and phototransduction in the eye. (Stryer, L. (1988) Biochemistry W.H. Freeman and Co., New York, pp. 975-980, 1029-1035.) Cytokines are produced in response to cell perturbation. Some cytokines are produced as precursor forms, and some form multimers in order to become active. They are produced

in groups and in patterns characteristic of the particular stimulus or disease, and the members of the group interact with one another and other molecules to produce an overall biological response. Interleukins, neurotrophins, growth factors, interferons, and chemokines are all families of cytokines which work in conjunction with cellular receptors to regulate cell proliferation and differentiation and to affect such activities, e.g., leukocyte migration and function, hematopoietic cell proliferation, temperature regulation, acute response to infections, tissue remodeling, and cell survival. Studies using antibodies or other drugs that modify the activity of a particular cytokine are used to elucidate the roles of individual cytokines in pathology and physiology.

Chemokines are a small chemoattractant cytokines which are active in leukocyte trafficking. Initially, chemokines were isolated and purified from inflamed tissues, but recently several chemokines have been discovered through molecular cloning techniques. Chemokines have been shown to be active in cell activation and migration, angiogenic and angiostatic activities, suppression of hematopoiesis, HIV infectivity, and promoting Th-1(IL-2-, interferon γ-stimulated) cytokine release.

Chemokines generally contain 70-100 amino acids and are subdivided into four subfamilies based on the presence and arrangement of conserved CXC, CC, CX3C and C motifs. The CXC (alpha), CC (beta), and CX3C chemokines contain four conserved cysteines. The CC subfamily is active on monocytes, lymphocytes, eosinophils, and mast cells; the CXC subfamily, on neutrophils; CX3C and C subfamilies, on T-cells. Many of the CC chemokines have been characterized functionally as well as structurally. (Callard, R. and Gearing, A. (1994) The Cytokine Facts Book, Academic Press, New York, NY, pp. 181-190, 210-213, 223-227.)

Growth and differentiation factors function in intercellular communication. Once secreted from the cell, some factors require oligomerization or association with ECM in order to function. Complex interactions among these factors and their receptors result in the stimulation or inhibition of cell division, cell differentiation, cell signaling, and cell motility. Some factors act on their cell of origin (autocrine signaling); on neighboring cells (paracrine signaling); or on distant cells (endocrine signaling).

There are three broad classes of growth and differentiation factors. The first class includes the large polypeptide growth factors, e.g., epidermal growth factor, fibroblast growth factor, transforming growth factor, insulin-like growth factor, and platelet-derived

growth factor. Each of these defines a family of related molecules which stimulate cell proliferation for wound healing, bone synthesis and remodeling, and regeneration of epithelial, epidermal, and connective tissues, and induce differentiation of embryonic tissues. Nerve growth factor functions specifically as a neurotrophic factor, and all induce differentiation of embryonic tissues. The second class includes the hematopoietic growth factors which stimulate the proliferation and differentiation of blood cells such as B-lymphocytes, T-lymphocytes, erythrocytes, platelets, eosinophils, basophils, neutrophils, macrophages, and their stem cell precursors. These factors include colony-stimulating factors, erythropoietin, and cytokines, e.g., interleukins, interferons (IFNs), and tumor necrosis factor (TNF). Cytokines are secreted by cells of the immune system and function in immunomodulation. The third class includes small peptide factors e.g., bombesin, vasopressin, oxytocin, endothelin, transferrin, angiotensin II, vasoactive intestinal peptide, and bradykinin, which function as hormones to regulate cellular functions other than proliferation.

Growth and differentiation factors have been shown to play critical roles in neoplastic transformation of cells in vitro and in tumor progression in vivo. Inappropriate expression of growth factors by tumor cells may contribute to vascularization and metastasis of melanotic tumors. In hematopoiesis, growth factor misregulation can result in anemias, leukemias and lymphomas. Certain growth factors, e.g., IFN, are cytotoxic to tumor cells both in vivo and in vitro. Moreover, growth factors and/or their receptors are related both structurally and functionally related to oncoproteins. In addition, growth factors affect transcriptional regulation of both proto-oncogenes and oncosuppressor genes. (Pimentel, E. (1994) Handbook of Growth Factors, CRC Press, Ann Arbor, MI, pp. 6-25.)

Proteolytic enzymes or proteases degrade proteins by reducing the activation energy needed for the hydrolysis of peptide bonds. The major families are the zinc, serine, cysteine, thiol, and carboxyl proteases.

Zinc proteases, e.g., carboxypeptidase A, have a zinc ion bound to the active site, recognize C-terminal residues that contain an aromatic or bulky aliphatic side chain, and hydrolyze the peptide bond adjacent to the C-terminal residues. Serine proteases have an active site serine residue and include digestive enzymes, e.g., trypsin and chymotrypsin, components of the complement and blood-clotting cascades, and enzymes that control the

degradation and turnover of extracellular matrix (ECM) molecules. Subfamilies of serine proteases include tryptases (cleavage after arginine or lysine), aspases (cleavage after aspartate), chymases (cleavage after phenylalanine or leucine), metases (cleavage after methionine), and serases (cleavage after serine). Cysteine proteases (e.g. cathepsin) are produced by monocytes, macrophages and other immune cells and are involved in diverse cellular processes ranging from the processing of precursor proteins to intracellular degradation. Overproduction of these enzymes can cause the tissue destruction associated with rheumatoid arthritis and asthma. Thiol proteases, e.g., papain, contain an active site cysteine and are widely distributed within tissues. Thiol proteases effect catalysis through a thiol ester intermediate facilitated by a proximal histidine side chain. Carboxyl proteases, e.g., pepsin, are active only under acidic conditions (pH 2 to 3). The active site of pepsin contains two aspartate residues; when one aspartate is ionized and the other is not, the enzyme is active. A common feature of the carboxyl proteases is that they are inhibited by very low concentrations (10^{-10} M) of the inhibitor pepstatin. A substrate analog which induces structural changes at the active site of a protease functions as an antagonist or inhibitor.

Guanosine triphosphate-binding proteins (G proteins) participate in intracellular signal transduction and control regulatory pathways through cell surface receptors. These receptors respond to hormones, growth factors, neuromodulators, or other signaling molecules, by binding GTP. Binding of GTP leads to the production of cAMP which controls phosphorylation and activation of other proteins. During this process, the hydrolysis of GTP acts as an energy source as well as an on-off switch for the GTPase activity.

The G proteins are small proteins which consist of single 21-30 kDa polypeptides. They can be classified into five subfamilies: Ras, Rho, Ran, Rab, and ADP-ribosylation factor. These proteins regulate cell growth, cell cycle control, protein secretion, and intracellular vesicle interaction. In particular, the Ras proteins are essential in transducing signals from receptor tyrosine kinases to serine/threonine kinases which control cell growth and differentiation. Mutant Ras proteins, which bind but can not hydrolyze GTP, are permanently activated and cause continuous cell proliferation or cancer.

All five subfamilies share common structural features and four conserved motifs, I to IV. Motif I is the most variable and has the signature of GXXXXGK, in which lysine

interacts with the β- and γ-phosphate groups of GTP. Motif II, III, and IV have DTAGQE, NKXD, and EXSAX as their respective signatures and regulate the binding of g-phosphate, GTP, and the guanine base of GTP, respectively. Most of the membrane-bound G proteins require a carboxy terminal isoprenyl group (CAAX), added posttranslationally, for membrane association and biological activity. The G proteins also have a variable effector region, located between motifs I and II, which is characterized as the interaction site for guanine nucleotide exchange factors or GTPase-activating proteins.

Eukaryotic cells are bound by a membrane and subdivided into membrane bound compartments. As membranes are impermeable to many ions and polar molecules, transport of these molecules is mediated by ion channels, ion pumps, transport proteins, or pumps. Symporters and antiporters regulate cytosolic pH by transporting ions and small molecules, e.g., amino acids, glucose, and drugs, across membranes; symporters transport small molecules and ions in the same direction, and antiporters, in the opposite direction. Transporter superfamilies include facilitative transporters and active ATP binding cassette transporters involved in multiple-drug resistance and the targeting of antigenic peptides to MHC Class I molecules. These transporters bind to a specific ion or other molecule and undergo conformational changes in order to transfer the ion or molecule across a membrane. Transport can occur by a passive, concentration-dependent mechanism or can be linked to an energy source such as ATP hydrolysis or an ion gradient.

Ion channels are formed by transmembrane proteins which form a lined passageway across the membrane through which water and ions, e.g., Na^+, K^+, Ca^{2+}, and Cl^-, enter and exit the cell. For example, chloride channels are involved in the regulation of the membrane electric potential as well as absorption and secretion of ions across the membrane. In intracellular membranes of the Golgi apparatus and endocytic vesicles, chloride channels also regulate organelle pH. Electrophysiological and pharmacological studies suggest that a variety of chloride channels exist in different cell types and that many of these channels have one or more protein kinase phosphorylation sites.

Ion pumps are ATPases which actively maintain membrane gradients. Ion pumps can be grouped into three classes, e.g., P, V, and F, according to their structure and function. All have one or more binding sites for ATP on the cytosolic face of the membrane. The P-class ion pumps consist of two α and two β transmembrane subunits, include Ca^{2+} ATPase and Na^+/K^+ ATPase, and function in transporting H^+, Na^+, K^+, and

Ca^{2+} ions. The V- and F-class ion pumps have similar structures, a cytosolic domain formed by at least five extrinsic polypeptides and at least 2 transmembrane proteins, and only transport H^+. F class H^+ pumps have been identified from the membranes of mitochondria and chloroplast, and V-class H^+ pumps regulate acidity inside lysosomes, endosomes, and plant vacuoles.

A family of structurally related intrinsic membrane proteins known as facilitative glucose transporters catalyze the movement of glucose and other selected sugars across the plasma membrane. The proteins in this family contain a highly conserved, large transmembrane domain made of 12 transmembrane α-helices, and several less conserved, asymmetric, cytoplasmic and exoplasmic domains. (Pessin, J. E., and Bell, G.I. (1992) Annu. Rev. Physiol. 54:911-930.)

Amino acid transport is mediated by Na^+ dependent amino acid transporters. These transporters are involved in gastrointestinal and renal uptake of dietary and cellular amino acids and the re-uptake of neurotransmitters. Transport of cationic amino acids is mediated by the system y+ family members and the cationic amino acid transporter (CAT) family. Members of the CAT family share a high degree of sequence homology, and each contains 12-14 putative transmembrane domains. (Ito, K. and Groudine, M. (1997) J. Biol. Chem. 272:26780-26786.)

Proton-coupled, 12 membrane-spanning domain transporters such as PEPT 1 and PEPT 2 are responsible for gastrointestinal absorption and for renal reabsorbtion of peptides using an electrochemical H^+ gradient as the driving force. A heterodimeric peptide transporter, consisting of TAP 1 and TAP 2, is associated with antigen processing. Peptide antigens are transported across the membrane of the endoplasmic reticulum so they can be presented to the major histocompatibility complex class I molecules. Each TAP protein consists of multiple hydrophobic membrane spanning segments and a highly conserved ATP-binding cassette. (Boll, M. et al. (1996) Proc. Natl. Acad. Sci. 93:284-289.)

Hormones are secreted molecules that circulate in the body fluids and bind to specific receptors on the surface of, or within, target tissue cells. Although they have diverse biochemical compositions and mechanisms of action, hormones can be grouped into two categories. One category consists of small lipophilic molecules that diffuse through the plasma membrane of target cells, bind to cytosolic or nuclear receptors, and

form a complex alters gene expression. Examples of this category include retinoic acid, thyroxine, and the cholesterol derived steroid hormones, progesterone, estrogen, testosterone, cortisol, and aldosterone. These hormones have a long half-life, e.g., several hours to days, and long-term effects of their target cells. Their solubility in the blood may be increased by their association with carrier molecules. Within the target cell nucleus, hormone/receptor complexes bind to specific response elements in target gene regulatory regions.

A second category consists of hydrophilic hormones that function by binding to cell surface receptors and transducing the signal across the plasma membrane. Examples of this category include amino acid derivatives, such as catecholamines, e.g., epinephrine, norepinephrine, and histamine; peptide hormones, e.g., glucagon, insulin, gastrin, secretin, cholecystokinin, adrenocorticotropic hormone, follicle stimulating hormone, luteinizing hormone, thyroid stimulating hormone, parathormone, and vasopressin. Peptide hormones are synthesized as inactive forms and stored in secretory vesicles. These hormones are activated by protease cleavage before being released from the cell. Many hydrophilic hormones have a very short half-life and effect, e.g., seconds to hours, and are inactivated by proteases in the blood. (Lodish et al. (1995) Molecular Cell Biology, Scientific American Books Inc., New York, NY, pp. 856-864.)

Neuropeptides and vasomediators (NP/VM) comprise a large family of endogenous signaling molecules. Included in the family are neurotransmitters such as bombesin, neuropeptide Y, neurotensin, neuromedin N, melanocortins, opioids, e.g., enkephalins, endorphins and dynorphins, galanin, somatostatin, tachykinins, vasopressin, and vasoactive intestinal peptide, and circulatory system-borne signaling molecules, e.g., angiotensin, complement, calcitonin, endothelins, formyl-methionyl peptides, glucagon, cholecystokinin and gastrin. These proteins are synthesized as "pre-pro" molecules, and are activated and inactivated by proteolytic cleavage. NP/VMs can transduce signals directly, modulate the activity or release of other neurotransmitters and hormones, and act as catalytic enzymes in cascades. The effects of NP/VMs range from extremely brief or long-lasting (melanocortin-mediated changes in skin melanin). Regulatory molecules turn individual genes or groups of genes on and off in response to various inductive mechanisms of the cell or organism; act as transcription factors by determining whether or not transcription is initiated, enhanced, or repressed; and splice transcripts as dictated in a

particular cell or tissue. Although they interact with short stretches of DNA scattered throughout the entire genome, most gene expression is regulated near the site at which transcription starts or within the open reading frame of the gene being expressed. The regulated stretches of the DNA can be simple and interact with only a single protein, or they can require several proteins acting as part of a complex to regulate gene expression. The external features of the double helix which provide recognition sites are hydrogen bond donor and acceptor groups, hydrophobic patches, major and minor grooves, and regular, repeated stretches of sequences which cause distinct bends in the helix. The surface features of the regulatory molecule are complementary to those of the DNA.

Many of the transcription factors incorporate one of a set of DNA-binding structural motifs, each of which contains either α helices or ß sheets and binds to the major groove of DNA. Seven of the structural motifs common to transcription factors are helix-turn-helix, homeodomains, zinc finger, steroid receptor, ß sheets, leucine zipper, and helix-loop-helix. (Pabo, C.O. and R.T. Sauer (1992) Ann. Rev. Biochem. 61:1053-95.) Other domains of transcription factors may form crucial contacts with the DNA. In addition, accessory proteins provide important interactions which may convert a particular protein complex to an activator or a repressor or may prevent binding. (Alberts, B. et al. (1994) Molecular Biology of the Cell, Garland Publishing Co, New York, NY pp. 401-474.)

The discovery of new human signal peptide-containing proteins and the polynucleotides encoding these molecules satisfies a need in the art by providing new compositions which are useful in the diagnosis, treatment, and prevention of cancer and immunological disorders.

SUMMARY OF THE INVENTION

The invention features a substantially purified human signal peptide-containing protein (SIGP), having an amino acid sequence selected from the group consisting of SEQ ID NO:1 SEQ ID NO:2, SEQ ID NO:3, SEQ ID NO:4, SEQ ID NO:5, SEQ ID NO:6, SEQ ID NO:7, SEQ ID NO:8, SEQ ID NO:9, SEQ ID NO:10, SEQ ID NO:11, SEQ ID NO:12, SEQ ID NO:13, SEQ ID NO:14, and SEQ ID NO:15.

The invention further provides isolated and substantially purified polynucleotides encoding SIGP. In a particular aspect, the polynucleotide has a nucleic acid sequence selected from the group consisting of SEQ ID NO:16, SEQ ID NO:17, SEQ ID NO:18,

SEQ ID NO:19, SEQ ID NO:20, SEQ ID NO:21, SEQ ID NO:22, SEQ ID NO:23, SEQ ID NO:24, SEQ ID NO:25, SEQ ID NO:26, SEQ ID NO:27, SEQ ID NO:28, SEQ ID NO:29, and SEQ ID NO:30.

In addition, the invention provides a polynucleotide, or fragment thereof, which hybridizes to any of the polynucleotides encoding an SIGP selected from the group consisting of SEQ ID NO:1 SEQ ID NO:2, SEQ ID NO:3, SEQ ID NO:4, SEQ ID NO:5, SEQ ID NO:6, SEQ ID NO:7, SEQ ID NO:8, SEQ ID NO:9, SEQ ID NO:10, SEQ ID NO:11, SEQ ID NO:12, SEQ ID NO:13, SEQ ID NO:14, and SEQ ID NO:15. In another aspect, the invention provides a composition comprising isolated and purified polynucleotides selected from the group consisting of SEQ ID NO:16, SEQ ID NO:17, SEQ ID NO:18, SEQ ID NO:19, SEQ ID NO:20, SEQ ID NO:21, SEQ ID NO:22, SEQ ID NO:23, SEQ ID NO:24, SEQ ID NO:25, SEQ ID NO:26, SEQ ID NO:27, SEQ ID NO:28, SEQ ID NO:29, and SEQ ID NO:30, or a fragment thereof.

The invention further provides a polynucleotide comprising the complement, or fragments thereof, of any one of the polynucleotides encoding SIGP. In another aspect, the invention provides compositions comprising isolated and purified polynucleotides comprising the complement of SEQ ID NO:16, SEQ ID NO:17, SEQ ID NO:18, SEQ ID NO:19, SEQ ID NO:20, SEQ ID NO:21, SEQ ID NO:22, SEQ ID NO:23, SEQ ID NO:24, SEQ ID NO:25, SEQ ID NO:26, SEQ ID NO:27, SEQ ID NO:28, SEQ ID NO:29, and SEQ ID NO:30, or fragments thereof.

The present invention further provides an expression vector containing at least a fragment of any one of the polynucleotides selected from the group consisting of SEQ ID NO:16, SEQ ID NO:17, SEQ ID NO:18, SEQ ID NO:19, SEQ ID NO:20, SEQ ID NO:21, SEQ ID NO:22, SEQ ID NO:23, SEQ ID NO:24, SEQ ID NO:25, SEQ ID NO:26, SEQ ID NO:27, SEQ ID NO:28, SEQ ID NO:29, and SEQ ID NO:30. In yet another aspect, the expression vector containing the polynucleotide is contained within a host cell.

The invention also provides a method for producing a polypeptide or a fragment thereof, the method comprising the steps of: (a) culturing the host cell containing an expression vector containing at least a fragment of a polynucleotide encoding SIGP under conditions suitable for the expression of the polypeptide; and (b) recovering the polypeptide from the host cell culture.

The invention also provides a pharmaceutical composition comprising a

substantially purified SIGP in conjunction with a suitable pharmaceutical carrier.

The invention further includes a purified antibody which binds to SIGP, as well as a purified agonist and a purified antagonist of SIGP.

The invention also provides a method for treating or preventing a cancer associated with the decreased expression or activity of SIGP, the method comprising the step of administering to a subject in need of such treatment an effective amount of a pharmaceutical composition containing SIGP.

The invention also provides a method for treating or preventing a cancer associated with the increased expression or activity of SIGP, the method comprising the step of administering to a subject in need of such treatment an effective amount of an antagonist of SIGP.

The invention also provides a method for treating or preventing an immune response associated with the increased expression or activity of SIGP, the method comprising the step of administering to a subject in need of such treatment an effective amount of an antagonist of SIGP.

The invention also provides a method for detecting a nucleic acid sequence which encodes a human regulatory proteins in a biological sample, the method comprising the steps of: a) hybridizing a nucleic acid sequence of the biological sample to a polynucleotide sequence complementary to the polynucleotide encoding SIGP, thereby forming a hybridization complex; and b) detecting the hybridization complex, wherein the presence of the hybridization complex correlates with the presence of the nucleic acid sequence encoding the human regulatory protein in the biological sample.

The invention also provides a microarray containing at least a fragment of at least one of the polynucleotides encoding a polypeptide having an amino acid sequence selected from the group consisting of SEQ ID NO:1 SEQ ID NO:2, SEQ ID NO:3, SEQ ID NO:4, SEQ ID NO:5, SEQ ID NO:6, SEQ ID NO:7, SEQ ID NO:8, SEQ ID NO:9, SEQ ID NO:10, SEQ ID NO:11, SEQ ID NO:12, SEQ ID NO:13, SEQ ID NO:14, and SEQ ID NO:15.

The invention also provides a method for detecting the expression level of a nucleic acid encoding a human regulatory protein in a biological sample, the method comprising the steps of hybridizing the nucleic acid sequence of the biological sample to a complementary polynucleotide, thereby forming hybridization complex; and determining

expression of the nucleic acid sequence encoding a human regulatory protein in the biological sample by identifying the presence of the hybridization complex. In a preferred embodiment, prior to the hybridizing step, the nucleic acid sequences of the biological sample are amplified and labeled by the polymerase chain reaction.

DESCRIPTION OF THE INVENTION

Before the present proteins, nucleotide sequences, and methods are described, it is understood that this invention is not limited to the particular methodology, protocols, cell lines, vectors, and reagents described, as these may vary. It is also to be understood that the terminology used herein is for the purpose of describing particular embodiments only, and is not intended to limit the scope of the present invention which will be limited only by the appended claims.

It must be noted that as used herein and in the appended claims, the singular forms "a," "an," and "the" include plural reference unless the context clearly dictates otherwise. Thus, for example, a reference to "a host cell" includes a plurality of such host cells, and a reference to "an antibody" is a reference to one or more antibodies and equivalents thereof known to those skilled in the art, and so forth.

Unless defined otherwise, all technical and scientific terms used herein have the same meanings as commonly understood by one of ordinary skill in the art to which this invention belongs. Although any methods and materials similar or equivalent to those described herein can be used in the practice or testing of the present invention, the preferred methods, devices, and materials are now described. All publications mentioned herein are cited for the purpose of describing and disclosing the cell lines, vectors, and methodologies which are reported in the publications and which might be used in connection with the invention. Nothing herein is to be construed as an admission that the invention is not entitled to antedate such disclosure by virtue of prior invention.

DEFINITIONS

"SIGP," as used herein, refers to the amino acid sequences of substantially purified SIGP obtained from any species, particularly a mammalian species, including bovine, ovine, porcine, murine, equine, and preferably the human species, from any source, whether natural, synthetic, semi-synthetic, or recombinant.

The term "agonist," as used herein, refers to a molecule which, when bound to SIGP, increases or prolongs the duration of the effect of SIGP. Agonists may include proteins, nucleic acids, carbohydrates, or any other molecules which bind to and modulate the effect of SIGP.

An "allele" or an "allelic sequence," as these terms are used herein, is an alternative form of the gene encoding SIGP. Alleles may result from at least one mutation in the nucleic acid sequence and may result in altered mRNAs or in polypeptides whose structure or function may or may not be altered. Any given natural or recombinant gene may have none, one, or many allelic forms. Common mutational changes which give rise to alleles are generally ascribed to natural deletions, additions, or substitutions of nucleotides. Each of these types of changes may occur alone, or in combination with the others, one or more times in a given sequence.

"Altered" nucleic acid sequences encoding SIGP, as described herein, include those sequences with deletions, insertions, or substitutions of different nucleotides, resulting in a polynucleotide the same SIGP or a polypeptide with at least one functional characteristic of SIGP. Included within this definition are polymorphisms which may or may not be readily detectable using a particular oligonucleotide probe of the polynucleotide encoding SIGP, and improper or unexpected hybridization to alleles, with a locus other than the normal chromosomal locus for the polynucleotide sequence encoding SIGP. The encoded protein may also be "altered," and may contain deletions, insertions, or substitutions of amino acid residues which produce a silent change and result in a functionally equivalent SIGP. Deliberate amino acid substitutions may be made on the basis of similarity in polarity, charge, solubility, hydrophobicity, hydrophilicity, and/or the amphipathic nature of the residues, as long as the biological or immunological activity of SIGP is retained. For example, negatively charged amino acids may include aspartic acid and glutamic acid, positively charged amino acids may include lysine and arginine, and amino acids with uncharged polar head groups having similar hydrophilicity values may include leucine, isoleucine, and valine; glycine and alanine; asparagine and glutamine; serine and threonine; and phenylalanine and tyrosine.

The terms "amino acid" or "amino acid sequence," as used herein, refer to an oligopeptide, peptide, polypeptide, or protein sequence, or a fragment of any of these, and to naturally occurring or synthetic molecules. In this context, "fragments", "immunogenic

fragments", or "antigenic fragments" refer to fragments of SIGP which are preferably about 5 to about 15 amino acids in length and which retain some biological activity or immunological activity of SIGP. Where "amino acid sequence" is recited herein to refer to an amino acid sequence of a naturally occurring protein molecule, "amino acid sequence" and like terms are not meant to limit the amino acid sequence to the complete native amino acid sequence associated with the recited protein molecule.

"Amplification," as used herein, relates to the production of additional copies of a nucleic acid sequence. Amplification is generally carried out using polymerase chain reaction (PCR) technologies well known in the art. (See, e.g., Dieffenbach, C.W. and G.S. Dveksler (1995) PCR Primer, a Laboratory Manual, Cold Spring Harbor Press, Plainview, NY, pp.1-5.)

The term "antagonist," as it is used herein, refers to a molecule which, when bound to SIGP, decreases the amount or the duration of the effect of the biological or immunological activity of SIGP. Antagonists may include proteins, nucleic acids, carbohydrates, antibodies, or any other molecules which decrease the effect of SIGP.

As used herein, the term "antibody" refers to intact molecules as well as to fragments thereof, such as Fa, $F(ab')_2$, and Fv fragments, which are capable of binding the epitopic determinant. Antibodies that bind SIGP polypeptides can be prepared using intact polypeptides or using fragments containing small peptides of interest as the immunizing antigen. The polypeptide or oligopeptide used to immunize an animal (e.g., a mouse, a rat, or a rabbit) can be derived from the translation of RNA, or synthesized chemically, and can be conjugated to a carrier protein if desired. Commonly used carriers that are chemically coupled to peptides include bovine serum albumin, thyroglobulin, and keyhole limpet hemocyanin (KLH). The coupled peptide is then used to immunize the animal.

The term "antigenic determinant," as used herein, refers to that fragment of a molecule (i.e., an epitope) that makes contact with a particular antibody. When a protein or a fragment of a protein is used to immunize a host animal, numerous regions of the protein may induce the production of antibodies which bind specifically to antigenic determinants (given regions or three-dimensional structures on the protein). An antigenic determinant may compete with the intact antigen (i.e., the immunogen used to elicit the immune response) for binding to an antibody.

The term "antisense," as used herein, refers to any composition containing a

nucleic acid sequence which is complementary to a specific nucleic acid sequence. The term "antisense strand" is used in reference to a nucleic acid strand that is complementary to the "sense" strand. Antisense molecules may be produced by any method including synthesis or transcription. Once introduced into a cell, the complementary nucleotides combine with natural sequences produced by the cell to form duplexes and to block either transcription or translation. The designation "negative" can refer to the antisense strand, and the designation "positive" can refer to the sense strand.

As used herein, the term "biologically active," refers to a protein having structural, regulatory, or biochemical functions of a naturally occurring molecule. Likewise, "immunologically active" refers to the capability of the natural, recombinant, or synthetic SIGP, or of any oligopeptide thereof, to induce a specific immune response in appropriate animals or cells and to bind with specific antibodies.

The terms "complementary" or "complementarity," as used herein, refer to the natural binding of polynucleotides under permissive salt and temperature conditions by base pairing. For example, the sequence "A-G-T" binds to the complementary sequence "T-C-A." Complementarity between two single-stranded molecules may be "partial," such that only some of the nucleic acids bind, or it may be "complete," such that total complementarity exists between the single stranded molecules. The degree of complementarity between nucleic acid strands has significant effects on the efficiency and strength of the hybridization between the nucleic acid strands. This is of particular importance in amplification reactions, which depend upon binding between nucleic acids strands, and in the design and use of peptide nucleic acid (PNA) molecules.

A "composition comprising a given polynucleotide sequence" or a "composition comprising a given amino acid sequence," as these terms are used herein, refer broadly to any composition containing the given polynucleotide or amino acid sequence. The composition may comprise a dry formulation, an aqueous solution, or a sterile composition. Compositions comprising polynucleotides encoding SIGP, e.g., SEQ ID NO:16, SEQ ID NO:17, SEQ ID NO:18, SEQ ID NO:19, SEQ ID NO:20, SEQ ID NO:21, SEQ ID NO:22, SEQ ID NO:23, SEQ ID NO:24, SEQ ID NO:25, SEQ ID NO:26, SEQ ID NO:27, SEQ ID NO:28, SEQ ID NO:29, and SEQ ID 30, or fragments thereof, may be employed as hybridization probes. The probes may be stored in freeze-dried form and may be associated with a stabilizing agent such as a carbohydrate. In hybridizations, the

probe may be deployed in an aqueous solution containing salts (e.g., NaCl), detergents (e.g., SDS) and other components (e.g., Denhardt's solution, dry milk, salmon sperm DNA, etc.).

The phrase "consensus sequence," as used herein, refers to a nucleic acid sequence which has been resequenced to resolve uncalled bases, extended using XL-PCR™ (Perkin Elmer, Norwalk, CT) in the 5' and/or the 3' direction, and resequenced, or which has been assembled from the overlapping sequences of more than one Incyte Clone using a computer program for fragment assembly, such as the GELVIEW™ Fragment Assembly system (GCG, Madison, WI). Some sequences have been both extended and assembled to produce the consensus sequence .

As used herein, the term "correlates with expression of a polynucleotide" indicates that the detection of the presence of nucleic acids, the same or related to a nucleic acid sequence encoding SIGP, by northern analysis is indicative of the presence of nucleic acids encoding SIGP in a sample, and thereby correlates with expression of the transcript from the polynucleotide encoding SIGP.

The term "SIGP" refers to any or all of the human polypeptides, SIGP-1, SIGP-2, SIGP-3, SIGP-4, SIGP-5, SIGP-6, SIGP-7, SIGP-8, SIGP-9, SIGP-10, SIGP-11, SIGP-12, SIGP-13, SIGP-14, and SIGP-15.

A "deletion," as the term is used herein, refers to a change in the amino acid or nucleotide sequence that results in the absence of one or more amino acid residues or nucleotides.

The term "derivative," as used herein, refers to the chemical modification of SIGP, of a polynucleotide sequence encoding SIGP, or of a polynucleotide sequence complementary to a polynucleotide sequence encoding SIGP. Chemical modifications of a polynucleotide sequence can include, for example, replacement of hydrogen by an alkyl, acyl, or amino group. A derivative polynucleotide encodes a polypeptide which retains at least one biological or immunological function of the natural molecule. A derivative polypeptide is one modified by glycosylation, pegylation, or any similar process that retains at least one biological or immunological function of the polypeptide from which it was derived.

The term "homology," as used herein, refers to a degree of complementarity. There may be partial homology or complete homology. The word "identity" may

substitute for the word "homology." A partially complementary sequence that at least partially inhibits an identical sequence from hybridizing to a target nucleic acid is referred to as "substantially homologous." The inhibition of hybridization of the completely complementary sequence to the target sequence may be examined using a hybridization assay (Southern or northern blot, solution hybridization, and the like) under conditions of reduced stringency. A substantially homologous sequence or hybridization probe will compete for and inhibit the binding of a completely homologous sequence to the target sequence under conditions of reduced stringency. This is not to say that conditions of reduced stringency are such that non-specific binding is permitted, as reduced stringency conditions require that the binding of two sequences to one another be a specific (i.e., a selective) interaction. The absence of non-specific binding may be tested by the use of a second target sequence which lacks even a partial degree of complementarity (e.g., less than about 30% homology or identity). In the absence of non-specific binding, the substantially homologous sequence or probe will not hybridize to the second non-complementary target sequence.

The phrases "percent identity" or "% identity" refer to the percentage of sequence similarity found in a comparison of two or more amino acid or nucleic acid sequences. Percent identity can be determined electronically, e.g., by using the MegAlign program (Lasergene software package, DNASTAR, Inc., Madison WI). The MegAlign program can create alignments between two or more sequences according to different methods, e.g., the Clustal Method. (Higgins, D.G. and Sharp, P.M. (1988) Gene 73:237-244.) The Clustal algorithm groups sequences into clusters by examining the distances between all pairs. The clusters are aligned pairwise and then in groups. The percentage similarity between two amino acid sequences, e.g., sequence A and sequence B, is calculated by dividing the length of sequence A, minus the number of gap residues in sequence A, minus the number of gap residues in sequence B, into the sum of the residue matches between sequence A and sequence B, times one hundred. Gaps of low or of no homology between the two amino acid sequences are not included in determining percentage similarity. Percent identity between nucleic acid sequences can also be calculated by the Clustal Method, or by other methods known in the art, such as the Jotun Hein Method. (See, e.g., Hein, J. (1990) Methods in Enzymology 183:626-645.) Identity between sequences can also be determined by other methods known in the art, e.g., by varying hybridization

conditions.

"Human artificial chromosomes" (HACs), as described herein, are linear microchromosomes which may contain DNA sequences of about 6 kb to 10 Mb in size, and which contain all of the elements required for stable mitotic chromosome segregation and maintenance. (See, e.g., Harrington, J.J. et al. (1997) Nat Genet. 15:345-355.)

The term "humanized antibody," as used herein, refers to antibody molecules in which the amino acid sequence in the non-antigen binding regions has been altered so that the antibody more closely resembles a human antibody, and still retains its original binding ability.

"Hybridization," as the term is used herein, refers to any process by which a strand of nucleic acid binds with a complementary strand through base pairing.

As used herein, the term "hybridization complex" as used herein, refers to a complex formed between two nucleic acid sequences by virtue of the formation of hydrogen bonds between complementary bases. A hybridization complex may be formed in solution (e.g., C_0t or R_0t analysis) or formed between one nucleic acid sequence present in solution and another nucleic acid sequence immobilized on a solid support (e.g., paper, membranes, filters, chips, pins or glass slides, or any other appropriate substrate to which cells or their nucleic acids have been fixed).

The words "insertion" or "addition," as used herein, refer to changes in an amino acid or nucleotide sequence resulting in the addition of one or more amino acid residues or nucleotides, respectively, to the sequence found in the naturally occurring molecule.

"Immune response" can refer to conditions associated with inflammation, trauma, immune disorders, or infectious or genetic disease, etc. These conditions can be characterized by expression of various factors, e.g., cytokines, chemokines, and other signaling molecules, which may affect cellular and systemic defense systems.

The term "microarray," as used herein, refers to an array of distinct polynucleotides or oligonucleotides arrayed on a substrate, such as paper, nylon or any other type of membrane, filter, chip, glass slide, or any other suitable solid support.

The term "modulate," as it appears herein, refers to a change in the activity of SIGP. For example, modulation may cause an increase or a decrease in protein activity, binding characteristics, or any other biological, functional, or immunological properties of SIGP.

The phrases "nucleic acid" or "nucleic acid sequence," as used herein, refer to an oligonucleotide, nucleotide, polynucleotide, or any fragment thereof, to DNA or RNA of genomic or synthetic origin which may be single-stranded or double-stranded and may represent the sense or the antisense strand, to peptide nucleic acid (PNA), or to any DNA-like or RNA-like material. In this context, "fragments" refers to those nucleic acid sequences which are greater than about 60 nucleotides in length, and most preferably are at least about 100 nucleotides, at least about 1000 nucleotides, or at least about 10,000 nucleotides in length.

The terms "operably associated" or "operably linked," as used herein, refer to functionally related nucleic acid sequences. A promoter is operably associated or operably linked with a coding sequence if the promoter controls the transcription of the encoded polypeptide. While operably associated or operably linked nucleic acid sequences can be contiguous and in reading frame, certain genetic elements, e.g., repressor genes, are not contiguously linked to the encoded polypeptide but still bind to operator sequences that control expression of the polypeptide.

The term "oligonucleotide," as used herein, refers to a nucleic acid sequence of at least about 6 nucleotides to 60 nucleotides, preferably about 15 to 30 nucleotides, and most preferably about 20 to 25 nucleotides, which can be used in PCR amplification or in a hybridization assay or microarray. As used herein, the term "oligonucleotide" is substantially equivalent to the terms "amplimers," "primers," "oligomers," and "probes," as these terms are commonly defined in the art.

"Peptide nucleic acid" (PNA), as used herein, refers to an antisense molecule or anti-gene agent which comprises an oligonucleotide of at least about 5 nucleotides in length linked to a peptide backbone of amino acid residues ending in lysine. The terminal lysine confers solubility to the composition. PNAs preferentially bind complementary single stranded DNA and RNA and stop transcript elongation, and may be pegylated to extend their lifespan in the cell. (See, e.g., Nielsen, P.E. et al. (1993) Anticancer Drug Des. 8:53-63.)

The term "sample," as used herein, is used in its broadest sense. A biological sample suspected of containing nucleic acids encoding SIGP, or fragments thereof, or SIGP itself may comprise a bodily fluid; an extract from a cell, chromosome, organelle, or membrane isolated from a cell; a cell; genomic DNA, RNA, or cDNA, in solution or

bound to a solid support; a tissue; a tissue print; etc.

As used herein, the terms "specific binding" or "specifically binding" refer to that interaction between a protein or peptide and an agonist, an antibody, or an antagonist. The interaction is dependent upon the presence of a particular structure of the protein recognized by the binding molecule (i.e., the antigenic determinant or epitope). For example, if an antibody is specific for epitope "A," the presence of a polypeptide containing the epitope A, or the presence of free unlabeled A, in a reaction containing free labeled A and the antibody will reduce the amount of labeled A that binds to the antibody.

As used herein, the term "stringent conditions" refers to conditions which permit hybridization between polynucleotide sequences and the claimed polynucleotide sequences. Suitably stringent conditions can be defined by, for example, the concentrations of salt or formamide in the prehybridization and hybridization solutions, or by the hybridization temperature, and are well known in the art. In particular, stringency can be increased by reducing the concentration of salt, increasing the concentration of formamide, or raising the hybridization temperature.

For example, hybridization under high stringency conditions could occur in about 50% formamide at about 37°C to 42°C. Hybridization could occur under reduced stringency conditions in about 35% to 25% formamide at about 30°C to 35°C. In particular, hybridization could occur under high stringency conditions at 42°C in 50% formamide, 5X SSPE, 0.3% SDS, and 200 μg/ml sheared and denatured salmon sperm DNA. Hybridization could occur under reduced stringency conditions as described above, but in 35% formamide at a reduced temperature of 35°C. The temperature range corresponding to a particular level of stringency can be further narrowed by calculating the purine to pyrimidine ratio of the nucleic acid of interest and adjusting the temperature accordingly. Variations on the above ranges and conditions are well known in the art.

The term "substantially purified," as used herein, refers to nucleic acid or amino acid sequences that are removed from their natural environment and are isolated or separated, and are at least about 60% free, preferably about 75% free, and most preferably about 90% free from other components with which they are naturally associated.

A "substitution," as used herein, refers to the replacement of one or more amino acids or nucleotides by different amino acids or nucleotides, respectively.

"Transformation," as defined herein, describes a process by which exogenous DNA

enters and changes a recipient cell. Transformation may occur under natural or artificial conditions according to various methods well known in the art, and may rely on any known method for the insertion of foreign nucleic acid sequences into a prokaryotic or eukaryotic host cell. The method for transformation is selected based on the type of host cell being transformed and may include, but is not limited to, viral infection, electroporation, heat shock, lipofection, and particle bombardment. The term "transformed" cells includes stably transformed cells in which the inserted DNA is capable of replication either as an autonomously replicating plasmid or as part of the host chromosome, and refers to cells which transiently express the inserted DNA or RNA for limited periods of time.

A "variant" of SIGP, as used herein, refers to an amino acid sequence that is altered by one or more amino acids. The variant may have "conservative" changes, wherein a substituted amino acid has similar structural or chemical properties (e.g., replacement of leucine with isoleucine). More rarely, a variant may have "nonconservative" changes (e.g., replacement of glycine with tryptophan). Analogous minor variations may also include amino acid deletions or insertions, or both. Guidance in determining which amino acid residues may be substituted, inserted, or deleted without abolishing biological or immunological activity may be found using computer programs well known in the art, for example, DNASTAR software.

THE INVENTION

The invention is based on the discovery of new human signal peptide-containing proteins, collectively referred to as SIGP and individually as SIGP-1, SIGP-2, SIGP-3, SIGP-4, SIGP-5, SIGP-6, SIGP-7, SIGP-8, SIGP-9, SIGP-10, SIGP-11, SIGP-12, SIGP-13, SIGP-14, and SIGP-15, the polynucleotides encoding SIGP (SEQ ID NO:16, SEQ ID NO:17, SEQ ID NO:18, SEQ ID NO:19, SEQ ID NO:20, SEQ ID NO:21, SEQ ID NO:22, SEQ ID NO:23, SEQ ID NO:24, SEQ ID NO:25, SEQ ID NO:26, SEQ ID NO:27, SEQ ID NO:28, SEQ ID NO:29, and SEQ ID NO:30); and the use of these compositions for the diagnosis, treatment, or prevention of cancer and immunological disorders. Table 1 shows the sequence identification numbers, Incyte Clone identification number, cDNA library, NCBI sequence identifier and GenBank species description for each of the human signal peptide-containing proteins disclosed herein.

TABLE 1

PROTEIN	NUCLEOTIDE	CLONE ID	LIBRARY	NCBI SEQ ID	HOMOLOG
SEQ ID NO:1	SEQ ID NO:16	866885	BRAITUT03	GI 1488683	Rattus norvegicus
SEQ ID NO:2	SEQ ID NO:17	1273453	TESTTUT02		
SEQ ID NO:3	SEQ ID NO:18	1534876	SPLNNOT04		
SEQ ID NO:4	SEQ ID NO:19	1634813	COLNNOT19	GI 219624	Mus musculus
SEQ ID NO:5	SEQ ID NO:20	1711840	PROSNOT16	GI 182651	Homo sapiens
SEQ ID NO:6	SEQ ID NO:21	1747327	STOMTUT02	GI 2062391	Homo sapien
SEQ ID NO:7	SEQ ID NO:22	1864292	PROSNOT19	GI 295671	Saccharomyces cerevisiae
SEQ ID NO:8	SEQ ID NO:23	1866437	THP1NOT01		
SEQ ID NO:9	SEQ ID NO:24	1871375	SKINBIT01		
SEQ ID NO:10	SEQ ID NO:25	1880830	LEUKNOT03	GI 1872521	Arabidopsis thaliana
SEQ ID NO:11	SEQ ID NO:26	2328134	COLNNOT11	GI 1911776	Homo sapiens
SEQ ID NO:12	SEQ ID NO:27	2652271	THYMNOT04	GI 895855	Solanum lycopersicum
SEQ ID NO:13	SEQ ID NO:28	2965248	SCORNOT04	GI 1478507	Homo sapiens
SEQ ID NO:14	SEQ ID NO:29	3057669	PONSAZT01	GI 260241	
SEQ ID NO:15	SEQ ID NO:30	3125156	LNODNOT05	GI 804750	Homo sapiens

Nucleic acids encoding the SIGP-1 of the present invention were first identified in Incyte Clone 866885 from the brain tumor cDNA library (BRAITUT03) using a computer search for amino acid sequence alignments. A consensus sequence, SEQ ID NO:16, was derived from Incyte Clones 866885 (BRAITUT03), 2991983 (KIDNFET02), 067954 (HUVESTB01), and 1499109 (SINTBST01).

In one embodiment, the invention encompasses a polypeptide comprising the amino acid sequence of SEQ ID NO:1. SIGP-1 is 236 amino acids in length and has a potential N-glycosylation site at N199; two potential casein kinase II phosphorylation sites at S8 and T72; a potential N-myristoylation site at G169; and three potential protein kinase C phosphorylation sites at T43, S96, and T201. SIGP-1 shares 24% identity with rat syntaxin (GI 1488683). The fragment of SEQ ID NO:16 from about nucleotide 43 to about nucleotide 93 is useful for hybridization. Northern analysis shows the expression of this sequence in hematopoietic and immune, reproductive, gastrointestinal, neural, cardiovascular, and developmental cDNA libraries. Approximately 43% of these libraries are associated with neoplastic disorders, 26% with inflammation, and 19% with cell proliferation.

Nucleic acids encoding the SIGP-2 of the present invention were first identified in Incyte Clone 1273453 from the testicle cDNA library (TESTTUT02) using a computer search for amino acid sequence alignments. A consensus sequence, SEQ ID NO:17, was derived from Incyte Clones 1273453 (TESTTUT02), 1970337 (UCMCL5T01), 1218926 (NEUTGMT01), 1881349 (LEUKNOT03), and 1722377 (BLADNT06).

In one embodiment, the invention encompasses a polypeptide comprising the amino acid sequence of SEQ ID NO:2. SIGP-17 is 267 amino acids in length and has a potential N glycosylation site at N230, five potential casein kinase II phosphorylation sites at S9, T45, T77, S190, and T263, and two potential protein kinase C phosphorylation sites at S232 and S236. The fragment of SEQ ID NO:17 from about nucleotide 140 to about nucleotide 175 is useful for hybridization. Northern analysis shows the expression of this sequence in reproductive, cardiovascular, and hematopoietic and immune cDNA libraries. Approximately 42% of these libraries are associated with neoplastic disorders and 40% with immune response.

Nucleic acids encoding the SIGP-3 of the present invention were first identified in

Incyte Clone 1534876 from the spleen cDNA library (SPLNNOT04) using a computer search for amino acid alignments. A consensus sequence, SEQ ID NO:18, was derived from Incyte Clones 1253004 (LUNGFET03), 1382838 (BRAITUT08), 1532501 (SPLNNOT04), 1534876 (SPLNNOT04), 1705806 (DUODNOT02), 1738301 (COLNNOT22), 1926209 (BRSTNOT02), and shotgun sequences SAOA00587, SAOA02048, and SAOA03535.

In one embodiment, the invention encompasses a polypeptide comprising the amino acid sequence of SEQ ID NO:3. SIGP-3 is 161 amino acids in length and has a potential signal peptide sequence between M1 and C13. SIGP-3 also has 17 cysteine residues with the potential for forming intramolecular disulfide bridges. Six of these cysteine residues, between residues C129 and C152, are found in a signature sequence for trypsin/alpha-amylase inhibitors that form a structure with intramolecular disulfide bridges. SIGP-3 has two potential casein kinase II phosphorylation sites at T25 and S35; and two potential protein kinase C phosphorylation sites at S35 and T87. The fragment of SEQ ID NO:18 from about nucleotide 406 to about nucleotide 477, which encompasses the trypsin/alpha-amylase inhibitor signature sequence, is useful for hybridization. Northern analysis shows the expression of this sequence in gastrointestinal and male and female reproductive cDNA libraries. Approximately 45% of these libraries are associated with neoplastic disorders and 28% with inflammation and the immune response..

Nucleic acids encoding the SIGP-4 of the present invention were first identified in Incyte Clone 1634813 from the cecal tissue cDNA library (COLNNOT19) using a computer search for amino acid sequence alignments. A consensus sequence, SEQ ID NO:19, was derived from Incyte Clones 1634813 (COLNNOT19), 2904583 (THYMNOT05), 1634813 (COLNNOT19), and 1310492 (COLNFET02), and shotgun sequence SAPA04436.

In one embodiment, the invention encompasses a polypeptide comprising the amino acid sequence of SEQ ID NO:4. SIGP-4 is 150 amino acids in length and has one potential N-glycosylation site at N139; and five potential phosphorylation sites at T48, S118, S126, S135, and S136. SIGP-4 also has a potential signal peptide sequence encompassing residues M1-A23. SIGP-4 shares 28% identity with mouse beta chemokine, Exodus-2 (GI 2196924). The fragment of SEQ ID NO:19 from about nucleotide 175 to about nucleotide 235 is useful for hybridization. Northern analysis

shows the expression of this sequence in gastrointestinal, developmental, hematopoietic, and immunological cDNA libraries. Approximately 50% of these libraries are associated with fetal development/cell proliferation and 25% with immune response.

Nucleic acids encoding the SIGP-5 of the present invention were first identified in Incyte Clone 1711840 from the prostate cDNA library (PROSNOT16) using a computer search for amino acid sequence alignments. A consensus sequence, SEQ ID NO:20, was derived from Incyte Clones 1711840 (PROSNOT16) and 2550483 (LUNGTUT06) and shotgun sequence SAQA03185.

In one embodiment, the invention encompasses a polypeptide comprising the amino acid sequence of SEQ ID NO:5. SIGP-5 is 118 amino acids in length and has three potential protein kinase C phosphorylation sites at S48, T103, and S109; and a potential signal peptide sequence from M1 to A20. SIGP-5 shares 61% identity with human midkine, a retinoic acid-responsive heparin binding factor involved in regulation of growth and differentiation (GI 182651). The fragment of SEQ ID NO:20 from about nucleotide 511 to about nucleotide 555 is useful for hybridization. Northern analysis shows the expression of this sequence in reproductive, gastrointestinal, developmental, neural, and cardiovascular cDNA libraries. Approximately 58% of these libraries are associated with cancer, 16% with immune response, and 23% with fetal/proliferating cells.

Nucleic acids encoding the SIGP-6 of the present invention were first identified in Incyte Clone 1747327 from the stomach tumor cDNA library (STOMTUT02) using a computer search for amino acid sequence alignments. A consensus sequence, SEQ ID NO:21, was derived from Incyte Clones 475228 (MMLR2DT01), 1500771 (SINTBST01), 1880656 (LEUKNOT03), 1747327 (STOMTUT02), and 2720285 (LUNGTUT10).

In one embodiment, the invention encompasses a polypeptide comprising the amino acid sequence of SEQ ID NO:6. SIGP-6 is 248 amino acids in length and has one potential N-glycosylation site at N56; three potential casein kinase II phosphorylation sites at S46, S134, and S140; and one potential protein kinase C phosphorylation site at T217. SIGP-6 shares 100% identity with human K12 protein precursor which is expressed in breast cancer cells and peripheral blood leukocytes (GI 2062391). Northern analysis shows the expression of this sequence in gastrointestinal,

reproductive, hematopoietic/immune, and cardiovascular cDNA libraries. Approximately 59% of these libraries are associated with cancer and 35% with immune response.

Nucleic acids encoding the SIGP-7 of the present invention were first identified in Incyte Clone 1864292 from the diseased prostate cDNA library (PROSNOT19) using a computer search for amino acid sequence alignments. A consensus sequence, SEQ ID NO:22, was derived from Incyte Clone 1864292 (PROSNOT19) and shotgun sequences SARA02195, SARA03070, SARA03675, and SATA02454.

In one embodiment, the invention encompasses a polypeptide comprising the amino acid sequence of SEQ ID NO:7. SIGP-7 is 404 amino acids in length and has one potential amidation site at V136; one potential cAMP- and cGMP-dependent protein kinase phosphorylation site at S66; twenty potential casein kinase II phosphorylation sites at S23, T27, T74, S110, S111, S118, T122, S143, S145, S205, S207, S218, S219, S220, T252, S254, S328, S330, S385, and T393; and twelve potential protein kinase C phosphorylation sites at T27, S76, T81, S140, S161, S176, S229, T285, S309, S356, S367, and S398. SIGP-7 shares 18% identity with the S. cerevisiae protein encoded by SRP40, a weak suppressor of a mutant of the subunit AC40 of DNA-dependent RNA polymerases I and II (GI 295671). The fragment of SEQ ID NO:22 f rom about nucleotide 193 to about nucleotide 222 is useful for hybridization. Northern analysis shows the expression of this sequence in reproductive, cardiovascular, and hematopoietic/immune cDNA libraries. Approximately 75% of these libraries are associated with cancer and 25% with immune response.

Nucleic acids encoding the SIGP-8 of the present invention were first identified in Incyte Clone 1866437 from the human promonocyte cell line cDNA library (THP1NOT01) using a computer search for amino acid sequence alignments. A consensus sequence, SEQ ID NO:23, was derived from Incyte Clones 817970 (OVARTUT01), 825684 (PROSNOT06), 1866437 (THP1NOT01), 2190170 (PROSNOT26), and 3137972 (SMCCNOT02).

In one embodiment, the invention encompasses a polypeptide comprising the amino acid sequence of SEQ ID NO:8. SIGP-8 is 405 amino acids in length and has one potential N-glycosylation site at N378; one potential cAMP- and cGMP-

phosphorylation site at S332; nine potential casein kinase II phosphorylation sites at T34, S51, T77, S107, S158, S264, T266, S296, and S332; and one potential protein kinase C phosphorylation site at S68. The fragment of SEQ ID NO:23 from about nucleotide 85 to about nucleotide 144 is useful for hybridization. Northern analysis shows the expression of this sequence in reproductive, hematopoietic/immune, neural, and developmental cDNA libraries. Approximately 37% of these libraries are associated with cancer, 33% with immune response, and 22% with fetal/proliferating cells.

Nucleic acids encoding the SIGP-9 of the present invention were first identified in Incyte Clone 1871375 from the leg skin erythema nodosum cDNA library (SKINBIT01) using a computer search for amino acid sequence alignments. A consensus sequence, SEQ ID NO:24, was derived from Incyte Clones 1428052 (SINTBST01), 1871375 (SKINBIT01), and 3210563 (BLADNOT08).

In one embodiment, the invention encompasses a polypeptide comprising the amino acid sequence of SEQ ID NO:9. SIGP-9 is 177 amino acids in length and has one potential casein kinase II phosphorylation site at S133; one potential glycosaminoglycan attachment site at S28GGG; and four potential protein kinase C phosphorylation sites at S44, S82, S115, and T148. SIGP-9 contains a signature sequence shared by the binding domains of receptors for lymphokines, hematopoietic growth factors and growth hormone-related molecules at S52RWSLWS. The fragment of SEQ ID NO:24 encoding the sequence surrounding the receptor binding domain signature from about nucleotide 190 to about nucleotide 249 is useful for hybridization. Northern analysis shows the expression of this sequence in reproductive, cardiovascular, gastrointestinal, and developmental cDNA libraries. Approximately 44% of these libraries are associated with cancer and 19% with immune response.

Nucleic acids encoding the SIGP-10 of the present invention were first identified in Incyte Clone 1880830 from the leukocyte cDNA library (LEUKNOT03) using a computer search for amino acid sequence alignments. A consensus sequence, SEQ ID NO:25, was derived from Incyte Clones 361577 (PROSNOT01); 2113591 (BRAITUT03); 1880830 (LEUKNOT03) and shotgun sequences SATA03292 and SATA00377.

In one embodiment, the invention encompasses a polypeptide comprising the

amino acid sequence of SEQ ID NO:10. SIGP-10 is 197 amino acids in length and has a potential cAMP- and cGMP-dependent protein kinase phosphorylation site at S121; and four potential protein kinase C phosphorylation sites at T3, S57, T107, and T153. SIGP-40 shares 15% identity with the Arabidopsis thaliana zinc-finger protein Lsd1 (GI 1872521). The fragment of SEQ ID NO:25 from about nucleotide 567 to about nucleotide 621 is useful for hybridization. Northern analysis shows the expression of this sequence in neural and reproductive cDNA libraries. Approximately 49% of these libraries are associated with neoplastic disorders, 24% with immune response, and 16% with fetal development.

Nucleic acids encoding the SIGP-11 of the present invention were first identified in Incyte Clone 2328134 from the colon cDNA library (COLNNOT11) using a computer search for amino acid sequence alignments. A consensus sequence, SEQ ID NO:26, was derived from Incyte Clones 2328134 (COLNNOT11), 1870180 (SKINBIT01), 081403 (SYNORAB01), and 851547 (NGANNOT01).

In one embodiment, the invention encompasses a polypeptide comprising the amino acid sequence of SEQ ID NO:11. SIGP-11 is 346 amino acids in length and has two potential cAMP- and cGMP-dependent protein kinase phosphorylation sites at residues S43 and S217; one potential casein kinase II phosphorylation site at residue T96; and five potential protein kinase C phosphorylation sites at residues T2, T15, T39, T247, and S301. SIGP-50 shares 33% identity with the human putative rab5-interacting protein (GI 1911776) and the casein kinase II phosphorylation site at residue T96. The fragment of SEQ ID NO:26 encoding the potential extracellular ligand binding domain from about nucleotide 16 to about nucleotide 76 is useful for hybridization. Northern analysis shows the expression of this sequence in reproductive, gastrointestinal, cardiovascular, and neural cDNA libraries. Approximately 44% of these libraries are associated with cancer, 28% are associated with immune response, and 20% with fetal disorders.

Nucleic acids encoding the SIGP-12 of the present invention were first identified in Incyte Clone 2652271 from the thymus cDNA library (THYMNOT04) using a computer search for amino acid sequence alignments. A consensus sequence, SEQ ID NO:27, was derived from Incyte Clones 2652271 (THYMNOT04), 2742813 (BRSTTUT14), 763431

(BRAITUT02), 1272403 (TESTTUT02), 1240531 (LUNGNOT03), and 1318448 (BLADNOT04).

In one embodiment, the invention encompasses a polypeptide comprising the amino acid sequence of SEQ ID NO:12. SIGP-12 is 256 amino acids in length and has three potential N glycosylation sites at N76, N106, and N212; three potential casein kinase II phosphorylation sites at T46, S188, and T204; two potential protein kinase C phosphorylation sites at S130 and S221; two potential ribonuclease T2 family histidine active sites from W62 to P69 and from F110 to C121; and a potential signal peptide sequence from M1 to A24. SIGP-59 shares 24% identity with Solanum lycopersicum ribonuclease LE (GI 895855); 80% identity between W62 and P75, one of the two ribonuclease T2 family histidine active sites; and 92% identity between F110 and C121, the second of the two ribonuclease T2 family histidine active sites. The fragment of SEQ ID NO:27 from about nucleotide 462 to about nucleotide 494 is useful for hybridization. Northern analysis shows the expression of this sequence in reproductive, hematopoietic, and gastrointestinal cDNA libraries. Approximately 53% of these libraries are associated with neoplastic disorders and 28% with immune response.

Nucleic acids encoding the SIGP-13 of the present invention were first identified in Incyte Clone 2965248 from the cervical spinal cord cDNA library (SCORNOT04) using a computer search for amino acid sequence alignments. A consensus sequence, SEQ ID NO:28, was derived from Incyte Clones 2965248 (SCORNOT04), 485746 (HNT2RAT01), 865684 (BRAITUT03), 1459157 (COLNFET02), 1597772 (BRAINOT14), 531430 (BRAINOT03), 725362 (SYNOOAT01), 1620429 (BRAITUT13), and 190305 (SYNORAB01).

In one embodiment, the invention encompasses a polypeptide comprising the amino acid sequence of SEQ ID NO:13. SIGP-13 is 235 amino acids in length and has seven potential cAMP- and cGMP-dependent protein kinase phosphorylation sites at S50, T80, T98, T126, S135, S136, and T194; three potential casein kinase II phosphorylation sites at S60, T80, and S81; six potential protein kinase C phosphorylation sites at S114, T119, T137, S142, S146, and S174; and a strathmin 1 family signature from P75 to E84. SIGP-28 shares 44% identity with human strathmin homolog SCG10/neuron-specific growth-associated protein in Alzheimer's disease (GI 1478503), and 71% identity between M1 and A107. In addition, one potential cAMP-

and cGMP-dependent protein kinase phosphorylation site, one potential casein kinase II phosphorylation site, the strathmin 1 family signature, and the hydrophobic transmembrane domains are conserved between these molecules. TM1 extends from about L15 to about F25; and TM2, from about G196 to about P212. The fragments of SEQ ID NO:28 from about nucleotide 158 to about nucleotide 196 and from about nucleotide 614 to about nucleotide 643 are useful for hybridization. Northern analysis shows the expression of this sequence in neural, reproductive, gastrointestinal, and hematopoietic/immune cDNA libraries. Approximately 50% of these libraries are associated with neoplastic disorders and 19% with immune response.

Nucleic acids encoding the SIGP-14 of the present invention were first identified in Incyte Clone 3057669 from the pons cDNA library (PONSAZT01) using a computer search for amino acid sequence alignments. A consensus sequence, SEQ ID NO:29, was derived from Incyte Clones 3057669 (PONSAZT01), 548211 (BEPINOT01), 3702516 (PENCNOT07), 3581270 (293TF3T01), 495191 (HNT2NOT01), 2784427 (BRSTNOT13), 1515961 (PANCTUT01), 3552333 (SYNONOT01), 2838668 (DRGLNOT01), 14600680 (COLNFET02), and 285677 (EOSIHET02).

In one embodiment, the invention encompasses a polypeptide comprising the amino acid sequence of SEQ ID NO:14. SIGP-14 is 371 amino acids in length and has three potential N-glycosylation sites at N70, N125, and N362; eleven potential casein kinase II phosphorylation sites at T22, S66, S72, S73, S102, T160, T201, T215, T278, T285, and S316; seven potential protein kinase C phosphorylation sites at S72, T79, S99, T127, S134, S257, and T299; and one protein kinase signature and profile from L188 to F200. Northern analysis shows the expression of this sequence in gastrointestinal, reproductive, and neural cDNA libraries. Approximately 54% of these libraries are associated with neoplastic disorders and 14% with immune response.

Nucleic acids encoding the SIGP-15 of the present invention were first identified in Incyte Clone 3125156 from the lymph node cDNA library (LNODNOT05) using a computer search for amino acid sequence alignments. A consensus sequence, SEQ ID NO:30, was derived from Incyte Clones 3125156 (LNODNOT05), 1417459 (BRAINOT12), 1567861 (UTRSNOT05), 154233 (THP1PLB02), 872652 (LUNGAST01), 2525803 (BRAITUT21), and 1209172 (BRSTNOT02).

In one embodiment, the invention encompasses a polypeptide comprising the

amino acid sequence of SEQ ID NO:15. SIGP-15 is 523 amino acids in length and has one potential N glycosylation sites at N186; nine potential casein kinase II phosphorylation sites at S63, T85, S179, S188, T210, S231, T269, T295, and S474; one potential glycosaminoglycan attachment site at S335; ten potential protein kinase C phosphorylation sites at T9, S159, S172, S179, T246, S263, S283, S416, S447, and S498; two potential tyrosine kinase phosphorylation sites at Y106 and Y170; and one tyrosine specific protein phosphatase active site at V331. SIGP-30 shares 21% identity with human T-cell protein tyrosine phosphatase (GI 804750), the N186 glycosylation site, the phosphorylation sites at S179, S188, T210, T246, S263, T295, S416, and Y170; and 50% identity between P324 and F344, the region of the tyrosine specific protein phosphatase active site. The fragments of SEQ ID NO:30 from about nucleotide 64 to about nucleotide 183 and from about nucleotide 1087 to about nucleotide 1119 are useful for hybridization. Northern analysis shows the expression of this sequence in neural, reproductive, and gastrointestinal cDNA libraries. Approximately 55% of these libraries are associated with neoplastic disorders and 22% with immune response.

The invention also encompasses SIGP variants. A preferred SIGP variant is one which has at least about 80%, more preferably at least about 90%, and most preferably at least about 95% amino acid sequence identity to the SIGP amino acid sequence, and which contains at least one functional or structural characteristic of SIGP.

The invention also encompasses polynucleotides which encode SIGP. Accordingly, any nucleic acid sequence which encodes the amino acid sequence of SIGP can be used to produce recombinant molecules which express SIGP. In a particular embodiment, the invention encompasses a polynucleotide consisting of a nucleic acid sequence selected from the group consisting of SEQ ID NO:16, SEQ ID NO:17, SEQ ID NO:18, SEQ ID NO:19, SEQ ID NO:20, SEQ ID NO:21, SEQ ID NO:22, SEQ ID NO:23, SEQ ID NO:24, SEQ ID NO:25, SEQ ID NO:26, SEQ ID NO:27, SEQ ID NO:28, SEQ ID NO:29, and SEQ ID NO:30.

It will be appreciated by those skilled in the art that as a result of the degeneracy of the genetic code, a multitude of polynucleotide sequences encoding SIGP, some bearing minimal homology to the polynucleotide sequences of any known and naturally occurring gene, may be produced. Thus, the invention contemplates each and every possible variation of polynucleotide sequence that could be made by selecting combinations based

on possible codon choices. These combinations are made in accordance with the standard triplet genetic code as applied to the polynucleotide sequence of naturally occurring SIGP, and all such variations are to be considered as being specifically disclosed.

Although nucleotide sequences which encode SIGP and its variants are preferably capable of hybridizing to the nucleotide sequence of the naturally occurring SIGP under appropriately selected conditions of stringency, it may be advantageous to produce nucleotide sequences encoding SIGP or its derivatives possessing a substantially different codon usage. Codons may be selected to increase the rate at which expression of the peptide occurs in a particular prokaryotic or eukaryotic host in accordance with the frequency with which particular codons are utilized by the host. Other reasons for substantially altering the nucleotide sequence encoding SIGP and its derivatives without altering the encoded amino acid sequences include the production of RNA transcripts having more desirable properties, such as a greater half-life, than transcripts produced from the naturally occurring sequence.

The invention also encompasses production of DNA sequences which encode SIGP and SIGP derivatives, or fragments thereof, entirely by synthetic chemistry. After production, the synthetic sequence may be inserted into any of the many available expression vectors and cell systems using reagents that are well known in the art. Moreover, synthetic chemistry may be used to introduce mutations into a sequence encoding SIGP or any fragment thereof.

Also encompassed by the invention are polynucleotide sequences that are capable of hybridizing to the claimed polynucleotide sequences, and, in particular, to those shown in SEQ ID NO:16, SEQ ID NO:17, SEQ ID NO:18, SEQ ID NO:19, SEQ ID NO:20, SEQ ID NO:21, SEQ ID NO:22, SEQ ID NO:23, SEQ ID NO:24, SEQ ID NO:25, SEQ ID NO:26, SEQ ID NO:27, SEQ ID NO:28, SEQ ID NO:29, and SEQ ID NO:30, under various conditions of stringency. (See, e.g., Wahl, G.M. and S.L. Berger (1987) Methods Enzymol. 152:399-407; and Kimmel, A.R. (1987) Methods Enzymol. 152:507-511.)

Methods for DNA sequencing are well known and generally available in the art and may be used to practice any of the embodiments of the invention. The methods may employ such enzymes as the Klenow fragment of DNA polymerase I, Sequenase® (US Biochemical Corp., Cleveland, OH), Taq polymerase (Perkin Elmer), thermostable T7 polymerase (Amersham, Chicago, IL), or combinations of polymerases and proofreading

exonucleases such as those found in the ELONGASE Amplification System (GIBCO/BRL, Gaithersburg, MD). Preferably, the process is automated with machines such as the Hamilton Micro Lab 2200 (Hamilton, Reno, NV), Peltier Thermal Cycler (PTC200; MJ Research, Watertown, MA) and the ABI Catalyst and 373 and 377 DNA Sequencers (Perkin Elmer).

The nucleic acid sequences encoding SIGP may be extended utilizing a partial nucleotide sequence and employing various methods known in the art to detect upstream sequences, such as promoters and regulatory elements. For example, one method which may be employed, restriction-site PCR, uses universal primers to retrieve unknown sequence adjacent to a known locus. (See, e.g., Sarkar, G. (1993) PCR Methods Applic. 2:318-322.) In particular, genomic DNA is first amplified in the presence of a primer complementary to a linker sequence within the vector and a primer specific to the region predicted to encode the gene. The amplified sequences are then subjected to a second round of PCR with the same linker primer and another specific primer internal to the first one. Products of each round of PCR are transcribed with an appropriate RNA polymerase and sequenced using reverse transcriptase.

Inverse PCR may also be used to amplify or extend sequences using divergent primers based on a known region. (See, e.g., Triglia, T. et al. (1988) Nucleic Acids Res. 16:8186.) The primers may be designed using commercially available software such as OLIGO 4.06 Primer Analysis software (National Biosciences Inc., Plymouth, MN) or another appropriate program to be about 22 to 30 nucleotides in length, to have a GC content of about 50% or more, and to anneal to the target sequence at temperatures of about 68°C to 72°C. The method uses several restriction enzymes to generate a suitable fragment in the known region of a gene. The fragment is then circularized by intramolecular ligation and used as a PCR template.

Another method which may be used is capture PCR, which involves PCR amplification of DNA fragments adjacent to a known sequence in human and yeast artificial chromosome DNA. (See, e.g., Lagerstrom, M. et al. (1991) PCR Methods Applic. 1:111-119.) In this method, multiple restriction enzyme digestions and ligations may be used to place an engineered double-stranded sequence into an unknown fragment of the DNA molecule before performing PCR. Other methods which may be used to retrieve unknown sequences are known in the art. (See, e.g., Parker, J.D. et al. (1991)

Nucleic Acids Res. 19:3055-3060.) Additionally, one may use PCR, nested primers, and PromoterFinder™ libraries to walk genomic DNA (Clontech, Palo Alto, CA). This process avoids the need to screen libraries and is useful in finding intron/exon junctions.

When screening for full-length cDNAs, it is preferable to use libraries that have been size-selected to include larger cDNAs. Also, random-primed libraries are preferable in that they will include more sequences which contain the 5' regions of genes. Use of a randomly primed library may be especially preferable for situations in which an oligo d(T) library does not yield a full-length cDNA. Genomic libraries may be useful for extension of sequence into 5' non-transcribed regulatory regions.

Capillary electrophoresis systems which are commercially available may be used to analyze the size or confirm the nucleotide sequence of sequencing or PCR products. In particular, capillary sequencing may employ flowable polymers for electrophoretic separation, four different fluorescent dyes (one for each nucleotide) which are laser activated, and a charge coupled device camera for detection of the emitted wavelengths. Output/light intensity may be converted to electrical signal using appropriate software (e.g., Genotyper™ and Sequence Navigator™, Perkin Elmer), and the entire process from loading of samples to computer analysis and electronic data display may be computer controlled. Capillary electrophoresis is especially preferable for the sequencing of small pieces of DNA which might be present in limited amounts in a particular sample.

In another embodiment of the invention, polynucleotide sequences or fragments thereof which encode SIGP may be used in recombinant DNA molecules to direct expression of SIGP, or fragments or functional equivalents thereof, in appropriate host cells. Due to the inherent degeneracy of the genetic code, other DNA sequences which encode substantially the same or a functionally equivalent amino acid sequence may be produced, and these sequences may be used to clone and express SIGP.

As will be understood by those of skill in the art, it may be advantageous to produce SIGP-encoding nucleotide sequences possessing non-naturally occurring codons. For example, codons preferred by a particular prokaryotic or eukaryotic host can be selected to increase the rate of protein expression or to produce an RNA transcript having desirable properties, such as a half-life which is longer than that of a transcript generated from the naturally occurring sequence.

The nucleotide sequences of the present invention can be engineered using methods

generally known in the art in order to alter SIGP-encoding sequences for a variety of reasons including, but not limited to, alterations which modify the cloning, processing, and/or expression of the gene product. DNA shuffling by random fragmentation and PCR reassembly of gene fragments and synthetic oligonucleotides may be used to engineer the nucleotide sequences. For example, site-directed mutagenesis may be used to insert new restriction sites, alter glycosylation patterns, change codon preference, produce splice variants, introduce mutations, and so forth.

In another embodiment of the invention, natural, modified, or recombinant nucleic acid sequences encoding SIGP may be ligated to a heterologous sequence to encode a fusion protein. For example, to screen peptide libraries for inhibitors of SIGP activity, it may be useful to encode a chimeric SIGP protein that can be recognized by a commercially available antibody. A fusion protein may also be engineered to contain a cleavage site located between the SIGP encoding sequence and the heterologous protein sequence, so that SIGP may be cleaved and purified away from the heterologous moiety.

In another embodiment, sequences encoding SIGP may be synthesized, in whole or in part, using chemical methods well known in the art. (See, e.g., Caruthers, M.H. et al. (1980) Nucl. Acids Res. Symp. Ser. 215-223, and Horn, T. et al. (1980) Nucl. Acids Res. Symp. Ser. 225-232.) Alternatively, the protein itself may be produced using chemical methods to synthesize the amino acid sequence of SIGP, or a fragment thereof. For example, peptide synthesis can be performed using various solid-phase techniques. (See, e.g., Roberge, J.Y. et al. (1995) Science 269:202-204.) Automated synthesis may be achieved using the ABI 431A Peptide Synthesizer (Perkin Elmer).

The newly synthesized peptide may be substantially purified by preparative high performance liquid chromatography. (See, e.g, Chiez, R.M. and F.Z. Regnier (1990) Methods Enzymol. 182:392-421.) The composition of the synthetic peptides may be confirmed by amino acid analysis or by sequencing. (See, e.g., Creighton, T. (1983) Proteins, Structures and Molecular Properties, WH Freeman and Co., New York, NY.) Additionally, the amino acid sequence of SIGP, or any part thereof, may be altered during direct synthesis and/or combined with sequences from other proteins, or any part thereof, to produce a variant polypeptide.

In order to express a biologically active SIGP, the nucleotide sequences encoding SIGP or derivatives thereof may be inserted into appropriate expression vector, i.e., a

vector which contains the necessary elements for the transcription and translation of the inserted coding sequence.

Methods which are well known to those skilled in the art may be used to construct expression vectors containing sequences encoding SIGP and appropriate transcriptional and translational control elements. These methods include in vitro recombinant DNA techniques, synthetic techniques, and in vivo genetic recombination. (See, e.g., Sambrook, J. et al. (1989) Molecular Cloning, A Laboratory Manual, Cold Spring Harbor Press, Plainview, NY, ch. 4, 8, and 16-17; and Ausubel, F.M. et al. (1995, and periodic supplements) Current Protocols in Molecular Biology, John Wiley & Sons, New York, NY, ch. 9, 13, and 16.)

A variety of expression vector/host systems may be utilized to contain and express sequences encoding SIGP. These include, but are not limited to, microorganisms such as bacteria transformed with recombinant bacteriophage, plasmid, or cosmid DNA expression vectors; yeast transformed with yeast expression vectors; insect cell systems infected with virus expression vectors (e.g., baculovirus); plant cell systems transformed with virus expression vectors (e.g., cauliflower mosaic virus (CaMV) or tobacco mosaic virus (TMV)) or with bacterial expression vectors (e.g., Ti or pBR322 plasmids); or animal cell systems.

The invention is not limited by the host cell employed.

The "control elements" or "regulatory sequences" are those non-translated regions, e.g., enhancers, promoters, and 5' and 3' untranslated regions, of the vector and polynucleotide sequences encoding SIGP which interact with host cellular proteins to carry out transcription and translation. Such elements may vary in their strength and specificity. Depending on the vector system and host utilized, any number of suitable transcription and translation elements, including constitutive and inducible promoters, may be used. For example, when cloning in bacterial systems, inducible promoters, e.g., hybrid lacZ promoter of the Bluescript® phagemid (Stratagene, La Jolla, CA) or pSport1™ plasmid (GIBCO/BRL), may be used. The baculovirus polyhedrin promoter may be used in insect cells. Promoters or enhancers derived from the genomes of plant cells (e.g., heat shock, RUBISCO, and storage protein genes) or from plant viruses (e.g., viral promoters or leader sequences) may be cloned into the vector. In mammalian cell systems, promoters from mammalian genes or from mammalian viruses are preferable. If it is

necessary to generate a cell line that contains multiple copies of the sequence encoding SIGP, vectors based on SV40 or EBV may be used with an appropriate selectable marker.

In bacterial systems, a number of expression vectors may be selected depending upon the use intended for SIGP. For example, when large quantities of SIGP are needed for the induction of antibodies, vectors which direct high level expression of fusion proteins that are readily purified may be used. Such vectors include, but are not limited to, multifunctional E. coli cloning and expression vectors such as Bluescript® (Stratagene), in which the sequence encoding SIGP may be ligated into the vector in frame with sequences for the amino-terminal Met and the subsequent 7 residues of ß-galactosidase so that a hybrid protein is produced, and pIN vectors. (See, e.g., Van Heeke, G. and S.M. Schuster (1989) J. Biol. Chem. 264:5503-5509.) pGEX vectors (Pharmacia Biotech, Uppsala, Sweden) may also be used to express foreign polypeptides as fusion proteins with glutathione S-transferase (GST). In general, such fusion proteins are soluble and can easily be purified from lysed cells by adsorption to glutathione-agarose beads followed by elution in the presence of free glutathione. Proteins made in such systems may be designed to include heparin, thrombin, or factor XA protease cleavage sites so that the cloned polypeptide of interest can be released from the GST moiety at will.

In the yeast Saccharomyces cerevisiae, a number of vectors containing constitutive or inducible promoters, such as alpha factor, alcohol oxidase, and PGH, may be used. (See, e.g., Ausubel, supra; and Grant et al. (1987) Methods Enzymol. 153:516-544.)

In cases where plant expression vectors are used, the expression of sequences encoding SIGP may be driven by any of a number of promoters. For example, viral promoters such as the 35S and 19S promoters of CaMV may be used alone or in combination with the omega leader sequence from TMV. (Takamatsu, N. (1987) EMBO J. 6:307-311.) Alternatively, plant promoters such as the small subunit of RUBISCO or heat shock promoters may be used. (See, e.g., Coruzzi, G. et al. (1984) EMBO J. 3:1671-1680; Broglie, R. et al. (1984) Science 224:838-843; and Winter, J. et al. (1991) Results Probl. Cell Differ. 17:85-105.) These constructs can be introduced into plant cells by direct DNA transformation or pathogen-mediated transfection. Such techniques are described in a number of generally available reviews. (See, e.g., Hobbs, S. or Murry, L.E. in McGraw Hill Yearbook of Science and Technology (1992) McGraw Hill, New York, NY; pp. 191-196.)

An insect system may also be used to express SIGP. For example, in one such system, Autographa californica nuclear polyhedrosis virus (AcNPV) is used as a vector to express foreign genes in Spodoptera frugiperda cells or in Trichoplusia larvae. The sequences encoding SIGP may be cloned into a non-essential region of the virus, such as the polyhedrin gene, and placed under control of the polyhedrin promoter. Successful insertion of sequences encoding SIGP will render the polyhedrin gene inactive and produce recombinant virus lacking coat protein. The recombinant viruses may then be used to infect, for example, S. frugiperda cells or Trichoplusia larvae in which SIGP may be expressed. (See, e.g., Engelhard, E.K. et al. (1994) Proc. Nat. Acad. Sci. 91:3224-3227.)

In mammalian host cells, a number of viral-based expression systems may be utilized. In cases where an adenovirus is used as an expression vector, sequences encoding SIGP may be ligated into an adenovirus transcription/translation complex consisting of the late promoter and tripartite leader sequence. Insertion in a non-essential E1 or E3 region of the viral genome may be used to obtain a viable virus which is capable of expressing SIGP in infected host cells. (See, e.g., Logan, J. and T. Shenk (1984) Proc. Natl. Acad. Sci. 81:3655-3659.) In addition, transcription enhancers, such as the Rous sarcoma virus (RSV) enhancer, may be used to increase expression in mammalian host cells.

Human artificial chromosomes (HACs) may also be employed to deliver larger fragments of DNA than can be contained and expressed in a plasmid. HACs of about 6 kb to 10 Mb are constructed and delivered via conventional delivery methods (liposomes, polycationic amino polymers, or vesicles) for therapeutic purposes.

Specific initiation signals may also be used to achieve more efficient translation of sequences encoding SIGP. Such signals include the ATG initiation codon and adjacent sequences. In cases where sequences encoding SIGP and its initiation codon and upstream sequences are inserted into the appropriate expression vector, no additional transcriptional or translational control signals may be needed. However, in cases where only coding sequence, or a fragment thereof, is inserted, exogenous translational control signals including the ATG initiation codon should be provided. Furthermore, the initiation codon should be in the correct reading frame to ensure translation of the entire insert. Exogenous translational elements and initiation codons may be of various origins, both natural and synthetic. The efficiency of expression may be enhanced by the inclusion of enhancers

appropriate for the particular cell system used. (See, e.g., Scharf, D. et al. (1994) Results Probl. Cell Differ. 20:125-162.)

In addition, a host cell strain may be chosen for its ability to modulate expression of the inserted sequences or to process the expressed protein in the desired fashion. Such modifications of the polypeptide include, but are not limited to, acetylation, carboxylation, glycosylation, phosphorylation, lipidation, and acylation. Post-translational processing which cleaves a "prepro" form of the protein may also be used to facilitate correct insertion, folding, and/or function. Different host cells which have specific cellular machinery and characteristic mechanisms for post-translational activities (e.g., CHO, HeLa, MDCK, HEK293, and WI38), are available from the American Type Culture Collection (ATCC, Bethesda, MD) and may be chosen to ensure the correct modification and processing of the foreign protein.

For long term, high yield production of recombinant proteins, stable expression is preferred. For example, cell lines capable of stably expressing SIGP can be transformed using expression vectors which may contain viral origins of replication and/or endogenous expression elements and a selectable marker gene on the same or on a separate vector. Following the introduction of the vector, cells may be allowed to grow for about 1 to 2 days in enriched media before being switched to selective media. The purpose of the selectable marker is to confer resistance to selection, and its presence allows growth and recovery of cells which successfully express the introduced sequences. Resistant clones of stably transformed cells may be proliferated using tissue culture techniques appropriate to the cell type.

Any number of selection systems may be used to recover transformed cell lines. These include, but are not limited to, the herpes simplex virus thymidine kinase genes and adenine phosphoribosyltransferase genes, which can be employed in *tk*$^-$ or *apr*$^-$ cells, respectively. (See, e.g., Wigler, M. et al. (1977) Cell 11:223-232; and Lowy, I. et al. (1980) Cell 22:817-823) Also, antimetabolite, antibiotic, or herbicide resistance can be used as the basis for selection. For example, *dhfr* confers resistance to methotrexate; *npt* confers resistance to the aminoglycosides neomycin and G-418; and *als* or *pat* confer resistance to chlorsulfuron and phosphinotricin acetyltransferase, respectively. (See, e.g., Wigler, M. et al. (1980) Proc. Natl. Acad. Sci. 77:3567-3570; Colbere-Garapin, F. et al (1981) J. Mol. Biol. 150:1-14; and Murry, supra.) Additional selectable genes have been

described, e.g., *trpB*, which allows cells to utilize indole in place of tryptophan, or *hisD*, which allows cells to utilize histinol in place of histidine. (See, e.g., Hartman, S.C. and R.C. Mulligan (1988) Proc. Natl. Acad. Sci. 85:8047-8051.) Recently, the use of visible markers has gained popularity with such markers as anthocyanins, ß glucuronidase and its substrate GUS, luciferase and its substrate luciferin. Green fluorescent proteins (GFP) (Clontech, Palo Alto, CA) are also used (See, e.g., Chalfie, M. et al. (1994) Science 263:802-805.) These markers can be used not only to identify transformants, but also to quantify the amount of transient or stable protein expression attributable to a specific vector system. (See, e.g., Rhodes, C.A. et al. (1995) Methods Mol. Biol. 55:121-131.)

Although the presence/absence of marker gene expression suggests that the gene of interest is also present, the presence and expression of the gene may need to be confirmed. For example, if the sequence encoding SIGP is inserted within a marker gene sequence, transformed cells containing sequences encoding SIGP can be identified by the absence of marker gene function. Alternatively, a marker gene can be placed in tandem with a sequence encoding SIGP under the control of a single promoter. Expression of the marker gene in response to induction or selection usually indicates expression of the tandem gene as well.

Alternatively, host cells which contain the nucleic acid sequence encoding SIGP and express SIGP may be identified by a variety of procedures known to those of skill in the art. These procedures include, but are not limited to, DNA-DNA or DNA-RNA hybridizations and protein bioassay or immunoassay techniques which include membrane, solution, or chip based technologies for the detection and/or quantification of nucleic acid or protein sequences.

The presence of polynucleotide sequences encoding SIGP can be detected by DNA-DNA or DNA-RNA hybridization or amplification using probes or fragments or fragments of polynucleotides encoding SIGP. Nucleic acid amplification based assays involve the use of oligonucleotides or oligomers based on the sequences encoding SIGP to detect transformants containing DNA or RNA encoding SIGP.

A variety of protocols for detecting and measuring the expression of SIGP, using either polyclonal or monoclonal antibodies specific for the protein, are known in the art. Examples of such techniques include enzyme-linked immunosorbent assays (ELISAs), radioimmunoassays (RIAs), and fluorescence activated cell sorting (FACS). A two-site,

monoclonal-based immunoassay utilizing monoclonal antibodies reactive to two non-interfering epitopes on SIGP is preferred, but a competitive binding assay may be employed. These and other assays are well described in the art. (See, e.g., Hampton, R. et al. (1990) Serological Methods, a Laboratory Manual, APS Press, St Paul, MN, Section IV; and Maddox, D.E. et al. (1983) J. Exp. Med. 158:1211-1216).

A wide variety of labels and conjugation techniques are known by those skilled in the art and may be used in various nucleic acid and amino acid assays. Means for producing labeled hybridization or PCR probes for detecting sequences related to polynucleotides encoding SIGP include oligolabeling, nick translation, end-labeling, or PCR amplification using a labeled nucleotide. Alternatively, the sequences encoding SIGP, or any fragments thereof, may be cloned into a vector for the production of an mRNA probe. Such vectors are known in the art, are commercially available, and may be used to synthesize RNA probes in vitro by addition of an appropriate RNA polymerase such as T7, T3, or SP6 and labeled nucleotides. These procedures may be conducted using a variety of commercially available kits, such as those provided by Pharmacia & Upjohn (Kalamazoo, MI), Promega (Madison, WI), and U.S. Biochemical Corp. (Cleveland, OH). Suitable reporter molecules or labels which may be used for ease of detection include radionuclides, enzymes, fluorescent, chemiluminescent, or chromogenic agents, as well as substrates, cofactors, inhibitors, magnetic particles, and the like.

Host cells transformed with nucleotide sequences encoding SIGP may be cultured under conditions suitable for the expression and recovery of the protein from cell culture. The protein produced by a transformed cell may be secreted or contained intracellularly depending on the sequence and/or the vector used. As will be understood by those of skill in the art, expression vectors containing polynucleotides which encode SIGP may be designed to contain signal sequences which direct secretion of SIGP through a prokaryotic or eukaryotic cell membrane. Other constructions may be used to join sequences encoding SIGP to nucleotide sequences encoding a polypeptide domain which will facilitate purification of soluble proteins. Such purification facilitating domains include, but are not limited to, metal chelating peptides such as histidine-tryptophan modules that allow purification on immobilized metals, protein A domains that allow purification on immobilized immunoglobulin, and the domain utilized in the FLAGS extension/affinity purification system (Immunex Corp., Seattle, WA). The inclusion of cleavable linker

sequences, such as those specific for Factor XA or enterokinase (Invitrogen, San Diego, CA), between the purification domain and the SIGP encoding sequence may be used to facilitate purification. One such expression vector provides for expression of a fusion protein containing SIGP and a nucleic acid encoding 6 histidine residues preceding a thioredoxin or an enterokinase cleavage site. The histidine residues facilitate purification on immobilized metal ion affinity chromatography. (IMAC) (See, e.g., Porath, J. et al. (1992) Prot. Exp. Purif. 3: 263-281.) The enterokinase cleavage site provides a means for purifying SIGP from the fusion protein. (See, e.g., Kroll, D.J. et al. (1993) DNA Cell Biol. 12:441-453.)

Fragments of SIGP may be produced not only by recombinant production, but also by direct peptide synthesis using solid-phase techniques. (See, e.g., Creighton, T.E. (1984) Protein: Structures and Molecular Properties, pp. 55-60, W.H. Freeman and Co., New York, NY.) Protein synthesis may be performed by manual techniques or by automation. Automated synthesis may be achieved, for example, using the Applied Biosystems 431A Peptide Synthesizer (Perkin Elmer). Various fragments of SIGP may be synthesized separately and then combined to produce the full length molecule.

THERAPEUTICS

The expression of the human signal peptide-containing proteins of the invention (SIGP) is closely associated with cell proliferation. Therefore, in cancers or immune response where SIGP is an activator, transcription factor, or enhancer, and is promoting cell proliferation, it is desirable to decrease the expression of SIGP. In conditions where SIGP is an inhibitor or suppressor and is controlling or decreasing cell proliferation, it is desirable to provide the protein or to increase the expression of SIGP.

In one embodiment, where SIGP is an inhibitor, SIGP or a fragment or derivative thereof may be administered to a subject to treat or prevent a cancer such as adenocarcinoma, leukemia, lymphoma, melanoma, myeloma, sarcoma, and teratocarcinoma. Such cancers include, but are not limited to, cancers of the adrenal gland, bladder, bone, bone marrow, brain, breast, cervix, gall bladder, ganglia, gastrointestinal tract, heart, kidney, liver, lung, muscle, ovary, pancreas, parathyroid, penis, prostate, salivary glands, skin, spleen, testis, thymus, thyroid, and uterus.

In another embodiment, a pharmaceutical composition comprising purified SIGP may

be used to treat or prevent a cancer including, but not limited to, those listed above.

In another embodiment, an agonist which is specific for SIGP may be administered to a subject to treat or prevent a cancer including, but not limited to, those cancers listed above.

In another further embodiment, a vector capable of expressing SIGP, or a fragment or a derivative thereof, may be administered to a subject to treat or prevent a cancer including, but not limited to, those cancers listed above.

In a further embodiment where SIGP is promoting cell proliferation, antagonists which decrease the expression or activity of SIGP may be administered to a subject to treat or prevent a cancer such as adenocarcinoma, leukemia, lymphoma, melanoma, myeloma, sarcoma, and teratocarcinoma. Such cancers include, but are not limited to, cancers of the adrenal gland, bladder, bone, bone marrow, brain, breast, cervix, gall bladder, ganglia, gastrointestinal tract, heart, kidney, liver, lung, muscle, ovary, pancreas, parathyroid, penis, prostate, salivary glands, skin, spleen, testis, thymus, thyroid, and uterus. In one aspect, antibodies which specifically bind SIGP may be used directly as an antagonist or indirectly as a targeting or delivery mechanism for bringing a pharmaceutical agent to cells or tissue which express SIGP.

In another embodiment, a vector expressing the complement of the polynucleotide encoding SIGP may be administered to a subject to treat or prevent a cancer including, but not limited to, those cancers listed above.

In yet another embodiment where SIGP is promoting leukocyte activity or proliferation, antagonists which decrease the activity of SIGP may be administered to a subject to treat or prevent an immune response. Such responses include, but are not limited to, disorders such as AIDS, Addison's disease, adult respiratory distress syndrome, allergies, anemia, asthma, atherosclerosis, bronchitis, cholecystitus, Crohn's disease, ulcerative colitis, atopic dermatitis, dermatomyositis, diabetes mellitus, emphysema, atrophic gastritis, glomerulonephritis, gout, Graves' disease, hypereosinophilia, irritable bowel syndrome, lupus erythematosus, multiple sclerosis, myasthenia gravis, myocardial or pericardial inflammation, osteoarthritis, osteoporosis, pancreatitis, polymyositis, rheumatoid arthritis, scleroderma, Sjögren's syndrome, and autoimmune thyroiditis; complications of cancer, hemodialysis, extracorporeal circulation; viral, bacterial, fungal, parasitic, protozoal, and helminthic infections; and trauma. In one aspect, antibodies

which specifically bind SIGP may be used directly as an antagonist or indirectly as a targeting or delivery mechanism for bringing a pharmaceutical agent to cells or tissue which express SIGP.

In another embodiment, a vector expressing the complement of the polynucleotide encoding SIGP may be administered to a subject to treat or prevent an immune response including, but not limited to, those listed above.

In other embodiments, any of the proteins, antagonists, antibodies, agonists, complementary sequences, or vectors of the invention may be administered in combination with other appropriate therapeutic agents. Selection of the appropriate agents for use in combination therapy may be made by one of ordinary skill in the art, according to conventional pharmaceutical principles. The combination of therapeutic agents may act synergistically to effect the treatment or prevention of the various disorders described above. Using this approach, one may be able to achieve therapeutic efficacy with lower dosages of each agent, thus reducing the potential for adverse side effects.

An antagonist of SIGP may be produced using methods which are generally known in the art. In particular, purified SIGP may be used to produce antibodies or to screen libraries of pharmaceutical agents to identify those which specifically bind SIGP. Antibodies to SIGP may also be generated using methods that are well known in the art. Such antibodies may include, but are not limited to, polyclonal, monoclonal, chimeric, and single chain antibodies, Fab fragments, and fragments produced by a Fab expression library. Neutralizing antibodies (i.e., those which inhibit dimer formation) are especially preferred for therapeutic use.

For the production of antibodies, various hosts including goats, rabbits, rats, mice, humans, and others may be immunized by injection with SIGP or with any fragment or oligopeptide thereof which has immunogenic properties. Depending on the host species, various adjuvants may be used to increase immunological response. Such adjuvants include, but are not limited to, Freund's, mineral gels such as aluminum hydroxide, and surface active substances such as lysolecithin, pluronic polyols, polyanions, peptides, oil emulsions, KLH, and dinitrophenol. Among adjuvants used in humans, BCG (bacilli Calmette-Guerin) and Corynebacterium parvum are especially preferable.

It is preferred that the oligopeptides, peptides, or fragments used to induce antibodies to SIGP have an amino acid sequence consisting of at least about 5 amino acids, and, more

preferably, of at least about 10 amino acids. It is also preferable that these oligopeptides, peptides, or fragments are identical to a portion of the amino acid sequence of the natural protein and contain the entire amino acid sequence of a small, naturally occurring molecule. Short stretches of SIGP amino acids may be fused with those of another protein, such as KLH, and antibodies to the chimeric molecule may be produced.

Monoclonal antibodies to SIGP may be prepared using any technique which provides for the production of antibody molecules by continuous cell lines in culture. These include, but are not limited to, the hybridoma technique, the human B-cell hybridoma technique, and the EBV-hybridoma technique. (See, e.g., Kohler, G. et al. (1975) Nature 256:495-497; Kozbor, D. et al. (1985) J. Immunol. Methods 81:31-42; Cote, R.J. et al. (1983) Proc. Natl. Acad. Sci. 80:2026-2030; and Cole, S.P. et al. (1984) Mol. Cell Biol. 62:109-120.)

In addition, techniques developed for the production of "chimeric antibodies," such as the splicing of mouse antibody genes to human antibody genes to obtain a molecule with appropriate antigen specificity and biological activity, can be used. (See, e.g., Morrison, S.L. et al. (1984) Proc. Natl. Acad. Sci. 81:6851-6855; Neuberger, M.S. et al. (1984) Nature 312:604-608; and Takeda, S. et al. (1985) Nature 314:452-454.) Alternatively, techniques described for the production of single chain antibodies may be adapted, using methods known in the art, to produce SIGP-specific single chain antibodies. Antibodies with related specificity, but of distinct idiotypic composition, may be generated by chain shuffling from random combinatorial immunoglobulin libraries. (See, e.g., Burton D.R. (1991) Proc. Natl. Acad. Sci. 88:10134-10137.)

Antibodies may also be produced by inducing in vivo production in the lymphocyte population or by screening immunoglobulin libraries or panels of highly specific binding reagents as disclosed in the literature. (See, e.g., Orlandi, R. et al. (1989) Proc. Natl. Acad. Sci. 86: 3833-3837; and Winter, G. et al. (1991) Nature 349:293-299.)

Antibody fragments which contain specific binding sites for SIGP may also be generated. For example, such fragments include, but are not limited to, F(ab')2 fragments produced by pepsin digestion of the antibody molecule and Fab fragments generated by reducing the disulfide bridges of the F(ab')2 fragments. Alternatively, Fab expression libraries may be constructed to allow rapid and easy identification of monoclonal Fab fragments with the desired specificity. (See, e.g., Huse, W.D. et al. (1989) Science

246:1275-1281.)

Various immunoassays may be used for screening to identify antibodies having the desired specificity. Numerous protocols for competitive binding or immunoradiometric assays using either polyclonal or monoclonal antibodies with established specificities are well known in the art. Such immunoassays typically involve the measurement of complex formation between SIGP and its specific antibody. A two-site, monoclonal-based immunoassay utilizing monoclonal antibodies reactive to two non-interfering SIGP epitopes is preferred, but a competitive binding assay may also be employed. (Maddox, supra.)

In another embodiment of the invention, the polynucleotides encoding SIGP, or any fragment or complement thereof, may be used for therapeutic purposes. In one aspect, the complement of the polynucleotide encoding SIGP may be used in situations in which it would be desirable to block the transcription of the mRNA. In particular, cells may be transformed with sequences complementary to polynucleotides encoding SIGP. Thus, complementary molecules or fragments may be used to modulate SIGP activity, or to achieve regulation of gene function. Such technology is now well known in the art, and sense or antisense oligonucleotides or larger fragments can be designed from various locations along the coding or control regions of sequences encoding SIGP.

Expression vectors derived from retroviruses, adenoviruses, or herpes or vaccinia viruses, or from various bacterial plasmids, may be used for delivery of nucleotide sequences to the targeted organ, tissue, or cell population. Methods which are well known to those skilled in the art can be used to construct vectors which will express nucleic acid sequences complementary to the polynucleotides of the gene encoding SIGP. (See, e.g., Sambrook, supra; and Ausubel, supra.)

Genes encoding SIGP can be turned off by transforming a cell or tissue with expression vectors which express high levels of a polynucleotide, or fragment thereof, encoding SIGP. Such constructs may be used to introduce untranslatable sense or antisense sequences into a cell. Even in the absence of integration into the DNA, such vectors may continue to transcribe RNA molecules until they are disabled by endogenous nucleases. Transient expression may last for a month or more with a non-replicating vector, and may last even longer if appropriate replication elements are part of the vector system.

As mentioned above, modifications of gene expression can be obtained by designing complementary sequences or antisense molecules (DNA, RNA, or PNA) to the control, 5', or regulatory regions of the gene encoding SIGP. Oligonucleotides derived from the transcription initiation site, e.g., between about positions -10 and +10 from the start site, are preferred. Similarly, inhibition can be achieved using triple helix base-pairing methodology. Triple helix pairing is useful because it causes inhibition of the ability of the double helix to open sufficiently for the binding of polymerases, transcription factors, or regulatory molecules. Recent therapeutic advances using triplex DNA have been described in the literature. (See, e.g., Gee, J.E. et al. (1994) in Huber, B.E. and B.I. Carr, Molecular and Immunologic Approaches, Futura Publishing Co., Mt. Kisco, NY, pp. 163-177.) A complementary sequence or antisense molecule may also be designed to block translation of mRNA by preventing the transcript from binding to ribosomes.

Ribozymes, enzymatic RNA molecules, may also be used to catalyze the specific cleavage of RNA. The mechanism of ribozyme action involves sequence-specific hybridization of the ribozyme molecule to complementary target RNA, followed by endonucleolytic cleavage. For example, engineered hammerhead motif ribozyme molecules may specifically and efficiently catalyze endonucleolytic cleavage of sequences encoding SIGP.

Specific ribozyme cleavage sites within any potential RNA target are initially identified by scanning the target molecule for ribozyme cleavage sites, including the following sequences: GUA, GUU, and GUC. Once identified, short RNA sequences of between 15 and 20 ribonucleotides, corresponding to the region of the target gene containing the cleavage site, may be evaluated for secondary structural features which may render the oligonucleotide inoperable. The suitability of candidate targets may also be evaluated by testing accessibility to hybridization with complementary oligonucleotides using ribonuclease protection assays.

Complementary ribonucleic acid molecules and ribozymes of the invention may be prepared by any method known in the art for the synthesis of nucleic acid molecules. These include techniques for chemically synthesizing oligonucleotides such as solid phase phosphoramidite chemical synthesis. Alternatively, RNA molecules may be generated by in vitro and in vivo transcription of DNA sequences encoding SIGP. Such DNA sequences may be incorporated into a wide variety of vectors with suitable RNA

polymerase promoters such as T7 or SP6. Alternatively, these cDNA constructs that synthesize complementary RNA, constitutively or inducibly, can be introduced into cell lines, cells, or tissues.

RNA molecules may be modified to increase intracellular stability and half-life. Possible modifications include, but are not limited to, the addition of flanking sequences at the 5' and/or 3' ends of the molecule, or the use of phosphorothioate or 2' O-methyl rather than phosphodiesterase linkages within the backbone of the molecule. This concept is inherent in the production of PNAs and can be extended in all of these molecules by the inclusion of nontraditional bases such as inosine, queosine, and wybutosine, as well as acetyl-, methyl-, thio-, and similarly modified forms of adenine, cytidine, guanine, thymine, and uridine which are not as easily recognized by endogenous endonucleases.

Many methods for introducing vectors into cells or tissues are available and equally suitable for use in vivo, in vitro, and ex vivo. For ex vivo therapy, vectors may be introduced into stem cells taken from the patient and clonally propagated for autologous transplant back into that same patient. Delivery by transfection, by liposome injections, or by polycationic amino polymers may be achieved using methods which are well known in the art. (See, e.g., Goldman, C.K. et al. (1997) Nature Biotechnology 15:462-466.)

Any of the therapeutic methods described above may be applied to any subject in need of such therapy, including, for example, mammals such as dogs, cats, cows, horses, rabbits, monkeys, and most preferably, humans.

An additional embodiment of the invention relates to the administration of a pharmaceutical or sterile composition, in conjunction with a pharmaceutically acceptable carrier, for any of the therapeutic effects discussed above. Such pharmaceutical compositions may consist of SIGP, antibodies to SIGP, and mimetics, agonists, antagonists, or inhibitors of SIGP. The compositions may be administered alone or in combination with at least one other agent, such as a stabilizing compound, which may be administered in any sterile, biocompatible pharmaceutical carrier including, but not limited to, saline, buffered saline, dextrose, and water. The compositions may be administered to a patient alone, or in combination with other agents, drugs, or hormones.

The pharmaceutical compositions utilized in this invention may be administered by any number of routes including, but not limited to, oral, intravenous, intramuscular, intra-arterial, intramedullary, intrathecal, intraventricular, transdermal, subcutaneous,

intraperitoneal, intranasal, enteral, topical, sublingual, or rectal means.

In addition to the active ingredients, these pharmaceutical compositions may contain suitable pharmaceutically-acceptable carriers comprising excipients and auxiliaries which facilitate processing of the active compounds into preparations which can be used pharmaceutically. Further details on techniques for formulation and administration may be found in the latest edition of Remington's Pharmaceutical Sciences (Maack Publishing Co., Easton, PA).

Pharmaceutical compositions for oral administration can be formulated using pharmaceutically acceptable carriers well known in the art in dosages suitable for oral administration. Such carriers enable the pharmaceutical compositions to be formulated as tablets, pills, dragees, capsules, liquids, gels, syrups, slurries, suspensions, and the like, for ingestion by the patient.

Pharmaceutical preparations for oral use can be obtained through combining active compounds with solid excipient and processing the resultant mixture of granules (optionally, after grinding) to obtain tablets or dragee cores. Suitable auxiliaries can be added, if desired. Suitable excipients include carbohydrate or protein fillers, such as sugars, including lactose, sucrose, mannitol, and sorbitol; starch from corn, wheat, rice, potato, or other plants; cellulose, such as methyl cellulose, hydroxypropylmethyl-cellulose, or sodium carboxymethylcellulose; gums, including arabic and tragacanth; and proteins, such as gelatin and collagen. If desired, disintegrating or solubilizing agents may be added, such as the cross-linked polyvinyl pyrrolidone, agar, and alginic acid or a salt thereof, such as sodium alginate.

Dragee cores may be used in conjunction with suitable coatings, such as concentrated sugar solutions, which may also contain gum arabic, talc, polyvinylpyrrolidone, carbopol gel, polyethylene glycol, and/or titanium dioxide, lacquer solutions, and suitable organic solvents or solvent mixtures. Dyestuffs or pigments may be added to the tablets or dragee coatings for product identification or to characterize the quantity of active compound, i.e., dosage.

Pharmaceutical preparations which can be used orally include push-fit capsules made of gelatin, as well as soft, sealed capsules made of gelatin and a coating, such as glycerol or sorbitol. Push-fit capsules can contain active ingredients mixed with fillers or binders, such as lactose or starches, lubricants, such as talc or magnesium stearate, and, optionally,

stabilizers. In soft capsules, the active compounds may be dissolved or suspended in suitable liquids, such as fatty oils, liquid, or liquid polyethylene glycol with or without stabilizers.

Pharmaceutical formulations suitable for parenteral administration may be formulated in aqueous solutions, preferably in physiologically compatible buffers such as Hanks's solution, Ringer's solution, or physiologically buffered saline. Aqueous injection suspensions may contain substances which increase the viscosity of the suspension, such as sodium carboxymethyl cellulose, sorbitol, or dextran. Additionally, suspensions of the active compounds may be prepared as appropriate oily injection suspensions. Suitable lipophilic solvents or vehicles include fatty oils, such as sesame oil, or synthetic fatty acid esters, such as ethyl oleate, triglycerides, or liposomes. Non-lipid polycationic amino polymers may also be used for delivery. Optionally, the suspension may also contain suitable stabilizers or agents to increase the solubility of the compounds and allow for the preparation of highly concentrated solutions.

For topical or nasal administration, penetrants appropriate to the particular barrier to be permeated are used in the formulation. Such penetrants are generally known in the art.

The pharmaceutical compositions of the present invention may be manufactured in a manner that is known in the art, e.g., by means of conventional mixing, dissolving, granulating, dragee-making, levigating, emulsifying, encapsulating, entrapping, or lyophilizing processes.

The pharmaceutical composition may be provided as a salt and can be formed with many acids, including but not limited to, hydrochloric, sulfuric, acetic, lactic, tartaric, malic, and succinic acid. Salts tend to be more soluble in aqueous or other protonic solvents than are the corresponding free base forms. In other cases, the preferred preparation may be a lyophilized powder which may contain any or all of the following: 1 mM to 50 mM histidine, 0.1% to 2% sucrose, and 2% to 7% mannitol, at a pH range of 4.5 to 5.5, that is combined with buffer prior to use.

After pharmaceutical compositions have been prepared, they can be placed in an appropriate container and labeled for treatment of an indicated condition. For administration of SIGP, such labeling would include amount, frequency, and method of administration.

Pharmaceutical compositions suitable for use in the invention include compositions

wherein the active ingredients are contained in an effective amount to achieve the intended purpose. The determination of an effective dose is well within the capability of those skilled in the art.

For any compound, the therapeutically effective dose can be estimated initially either in cell culture assays, e.g., of neoplastic cells or in animal models such as mice, rats, rabbits, dogs, or pigs. An animal model may also be used to determine the appropriate concentration range and route of administration. Such information can then be used to determine useful doses and routes for administration in humans.

A therapeutically effective dose refers to that amount of active ingredient, for example SIGP or fragments thereof, antibodies of SIGP, and agonists, antagonists or inhibitors of SIGP, which ameliorates the symptoms or condition. Therapeutic efficacy and toxicity may be determined by standard pharmaceutical procedures in cell cultures or with experimental animals, such as by calculating the ED50 (the dose therapeutically effective in 50% of the population) or LD50 (the dose lethal to 50% of the population) statistics. The dose ratio of therapeutic to toxic effects is the therapeutic index, and it can be expressed as the ED50/LD50 ratio. Pharmaceutical compositions which exhibit large therapeutic indices are preferred. The data obtained from cell culture assays and animal studies are used to formulate a range of dosage for human use. The dosage contained in such compositions is preferably within a range of circulating concentrations that includes the ED50 with little or no toxicity. The dosage varies within this range depending upon the dosage form employed, the sensitivity of the patient, and the route of administration.

The exact dosage will be determined by the practitioner, in light of factors related to the subject requiring treatment. Dosage and administration are adjusted to provide sufficient levels of the active moiety or to maintain the desired effect. Factors which may be taken into account include the severity of the disease state, the general health of the subject, the age, weight, and gender of the subject, time and frequency of administration, drug combination(s), reaction sensitivities, and response to therapy. Long-acting pharmaceutical compositions may be administered every 3 to 4 days, every week, or biweekly depending on the half-life and clearance rate of the particular formulation.

Normal dosage amounts may vary from about 0.1 μg to 100,000 μg, up to a total dose of about 1 gram, depending upon the route of administration. Guidance as to particular dosages and methods of delivery is provided in the literature and generally available to

practitioners in the art. Those skilled in the art will employ different formulations for nucleotides than for proteins or their inhibitors. Similarly, delivery of polynucleotides or polypeptides will be specific to particular cells, conditions, locations, etc.

DIAGNOSTICS

In another embodiment, antibodies which specifically bind SIGP may be used for the diagnosis of disorders characterized by expression of SIGP, or in assays to monitor patients being treated with SIGP or agonists, antagonists, or inhibitors of SIGP. Antibodies useful for diagnostic purposes may be prepared in the same manner as described above for therapeutics. Diagnostic assays for SIGP include methods which utilize the antibody and a label to detect SIGP in human body fluids or in extracts of cells or tissues. The antibodies may be used with or without modification, and may be labeled by covalent or non-covalent attachment of a reporter molecule. A wide variety of reporter molecules, several of which are described above, are known in the art and may be used.

A variety of protocols for measuring SIGP, including ELISAs, RIAs, and FACS, are known in the art and provide a basis for diagnosing altered or abnormal levels of SIGP expression. Normal or standard values for SIGP expression are established by combining body fluids or cell extracts taken from normal mammalian subjects, preferably human, with antibody to SIGP under conditions suitable for complex formation The amount of standard complex formation may be quantitated by various methods, preferably by photometric means. Quantities of SIGP expressed in subject, control, and disease samples from biopsied tissues are compared with the standard values. Deviation between standard and subject values establishes the parameters for diagnosing disease.

In another embodiment of the invention, the polynucleotides encoding SIGP may be used for diagnostic purposes. The polynucleotides which may be used include oligonucleotide sequences, complementary RNA and DNA molecules, and PNAs. The polynucleotides may be used to detect and quantitate gene expression in biopsied tissues in which expression of SIGP may be correlated with disease. The diagnostic assay may be used to determine absence, presence, and excess expression of SIGP, and to monitor regulation of SIGP levels during therapeutic intervention.

In one aspect, hybridization with PCR probes which are capable of detecting polynucleotide sequences, including genomic sequences, encoding SIGP or closely related

molecules may be used to identify nucleic acid sequences which encode SIGP. The specificity of the probe, whether it is made from a highly specific region, e.g., the 5' regulatory region, or from a less specific region, e.g., a conserved motif, and the stringency of the hybridization or amplification (maximal, high, intermediate, or low), will determine whether the probe identifies only naturally occurring sequences encoding SIGP, alleles, or related sequences.

Probes may also be used for the detection of related sequences, and should preferably contain at least 50% of the nucleotides from any of the SIGP encoding sequences. The hybridization probes of the subject invention may be DNA or RNA and may be derived from the sequence of SEQ ID NO:16, SEQ ID NO:17, SEQ ID NO:18, SEQ ID NO:19, SEQ ID NO:20, SEQ ID NO:21, SEQ ID NO:22, SEQ ID NO:23, SEQ ID NO:24, SEQ ID NO:25, SEQ ID NO:26, SEQ ID NO:27, SEQ ID NO:28, SEQ ID NO:29, and SEQ ID NO:30, or from genomic sequences including promoters, enhancers, and introns of the SIGP gene.

Means for producing specific hybridization probes for DNAs encoding SIGP include the cloning of polynucleotide sequences encoding SIGP or SIGP derivatives into vectors for the production of mRNA probes. Such vectors are known in the art, are commercially available, and may be used to synthesize RNA probes in vitro by means of the addition of the appropriate RNA polymerases and the appropriate labeled nucleotides. Hybridization probes may be labeled by a variety of reporter groups, for example, by radionuclides such as ^{32}P or ^{35}S, or by enzymatic labels, such as alkaline phosphatase coupled to the probe via avidin/biotin coupling systems, and the like.

Polynucleotide sequences encoding SIGP may be used for the diagnosis of a disorder associated with either increased or decreased expression of SIGP. Examples of such a disorder include, but are not limited to, cancers such as adenocarcinoma, leukemia, lymphoma, melanoma, myeloma, sarcoma, teratocarcinoma, and cancers of the adrenal gland, bladder, bone, brain, breast, cervix, gall bladder, ganglia, gastrointestinal tract, heart, kidney, liver, lung, bone marrow, muscle, ovary, pancreas, parathyroid, penis, prostate, salivary glands, skin, spleen, testis, thymus, thyroid, and uterus; neuronal disorders such as akathesia, Alzheimer's disease, amnesia, amyotrophic lateral sclerosis, bipolar disorder, catatonia, cerebral neoplasms, dementia, depression, Down's syndrome, tardive dyskinesia, dystonias, epilepsy, Huntington's disease, multiple sclerosis,

neurofibromatosis, Parkinson's disease, paranoid psychoses, schizophrenia, and Tourette's disorder; and immunological disorders such as AIDS, Addison's disease, adult respiratory distress syndrome, allergies, anemia, asthma, atherosclerosis, bronchitis, cholecystitus, Crohn's disease, ulcerative colitis, atopic dermatitis, dermatomyositis, diabetes mellitus, emphysema, atrophic gastritis, glomerulonephritis, gout, Graves' disease, hypereosinophilia, irritable bowel syndrome, lupus erythematosus, multiple sclerosis, myasthenia gravis, myocardial or pericardial inflammation, osteoarthritis, osteoporosis, pancreatitis, polymyositis, rheumatoid arthritis, scleroderma, Sjögren's syndrome, and thyroiditis. The polynucleotide sequences encoding SIGP may be used in Southern or northern analysis, dot blot, or other membrane-based technologies; in PCR technologies; in dipstick, pin, and ELISA assays; and in microarrays utilizing fluids or tissues from patients to detect altered SIGP expression. Such qualitative or quantitative methods are well known in the art.

In a particular aspect, the nucleotide sequences encoding SIGP may be useful in assays that detect the presence of associated disorders, particularly those mentioned above. The nucleotide sequences encoding SIGP may be labeled by standard methods and added to a fluid or tissue sample from a patient under conditions suitable for the formation of hybridization complexes. After a suitable incubation period, the sample is washed and the signal is quantitated and compared with a standard value. If the amount of signal in the patient sample is significantly altered in comparison to a control sample then the presence of altered levels of nucleotide sequences encoding SIGP in the sample indicates the presence of the associated disorder. Such assays may also be used to evaluate the efficacy of a particular therapeutic treatment regimen in animal studies, in clinical trials, or to monitor the treatment of an individual patient.

In order to provide a basis for the diagnosis of a disorder associated with expression of SIGP, a normal or standard profile for expression is established. This may be accomplished by combining body fluids or cell extracts taken from normal subjects, either animal or human, with a sequence, or a fragment thereof, encoding SIGP, under conditions suitable for hybridization or amplification. Standard hybridization may be quantified by comparing the values obtained from normal subjects with values from an experiment in which a known amount of a substantially purified polynucleotide is used. Standard values obtained in this manner may be compared with values obtained from samples from

patients who are symptomatic for a disorder. Deviation from standard values is used to establish the presence of a disorder.

Once the presence of a disorder is established and a treatment protocol is initiated, hybridization assays may be repeated on a regular basis to determine if the level of expression in the patient begins to approximate that which is observed in the normal subject. The results obtained from successive assays may be used to show the efficacy of treatment over a period ranging from several days to months.

With respect to cancer, the presence of a relatively high amount of transcript in biopsied tissue from an individual may indicate a predisposition for the development of the disease, or may provide a means for detecting the disease prior to the appearance of actual clinical symptoms. A more definitive diagnosis of this type may allow health professionals to employ preventative measures or aggressive treatment earlier thereby preventing the development or further progression of the cancer.

Additional diagnostic uses for oligonucleotides designed from the sequences encoding SIGP may involve the use of PCR. These oligomers may be chemically synthesized, generated enzymatically, or produced in vitro. Oligomers will preferably contain a fragment of a polynucleotide encoding SIGP, or a fragment of a polynucleotide complementary to the polynucleotide encoding SIGP, and will be employed under optimized conditions for identification of a specific gene or condition. Oligomers may also be employed under less stringent conditions for detection or quantitation of closely related DNA or RNA sequences.

Methods which may also be used to quantitate the expression of SIGP include radiolabeling or biotinylating nucleotides, coamplification of a control nucleic acid, and interpolating results from standard curves. (See, e.g., Melby, P.C. et al. (1993) J. Immunol. Methods 159:235-244; and Duplaa, C. et al. (1993) Anal. Biochem. 229-236.) The speed of quantitation of multiple samples may be accelerated by running the assay in an ELISA format where the oligomer of interest is presented in various dilutions and a spectrophotometric or colorimetric response gives rapid quantitation.

In further embodiments, oligonucleotides or longer fragments derived from any of the polynucleotide sequences described herein may be used as targets in a microarray. The microarray can be used to monitor the expression level of large numbers of genes simultaneously and to identify genetic variants, mutations, and polymorphisms. This

information may be used to determine gene function, to understand the genetic basis of a disorder, to diagnose a disorder, and to develop and monitor the activities of therapeutic agents.

In one embodiment, the microarray is prepared and used according to methods known in the art. (See, e.g., Chee et al. (1995) PCT application WO95/11995; Lockhart, D. J. et al. (1996) Nat. Biotech. 14:1675-1680; and Schena, M. et al. (1996) Proc. Natl. Acad. Sci. 93:10614-10619.)

The microarray is preferably composed of a large number of unique single-stranded nucleic acid sequences, usually either synthetic antisense oligonucleotides or fragments of cDNAs. The oligonucleotides are preferably about 6 to 60 nucleotides in length, more preferably about 15 to 30 nucleotides in length, and most preferably about 20 to 25 nucleotides in length. It may be preferable to use oligonucleotides which are about 7 to 10 nucleotides in length. The microarray may contain oligonucleotides which cover the known 5' or 3' sequence, sequential oligonucleotides which cover the full length sequence, or unique oligonucleotides selected from particular areas along the length of the sequence. Polynucleotides used in the microarray may be oligonucleotides specific to a gene or genes of interest. Oligonucleotides can also be specific to one or more unidentified cDNAs associated with a particular cell type or tissue type. It may be appropriate to use pairs of oligonucleotides on a microarray. The first oligonucleotide in each pair differs from the second oligonucleotide by one nucleotide. This nucleotide is preferably located in the center of the sequence. The second oligonucleotide serves as a control. The number of oligonucleotide pairs may range from about 2 to 1,000,000.

In order to produce oligonucleotides for use on a microarray, the gene of interest is examined using a computer algorithm which starts at the 5' end, or, more preferably, at the 3' end of the nucleotide sequence. The algorithm identifies oligomers of defined length that are unique to the gene, have a GC content within a range suitable for hybridization, and lack secondary structure that may interfere with hybridization. In one aspect, the oligomers may be synthesized on a substrate using a light-directed chemical process. (See, e.g., Chee et al., supra.) The substrate may be any suitable solid support, e.g., paper, nylon, any other type of membrane, or a filter, chip, or glass slide.

In another aspect, the oligonucleotides may be synthesized on the surface of the substrate using a chemical coupling procedure and an ink jet application apparatus. (See,

e.g., Baldeschweiler et al. (1995) PCT application WO95/251116.) An array analogous to a dot or slot blot (HYBRIDOT® apparatus, GIBCO/BRL) may be used to arrange and link cDNA fragments or oligonucleotides to the surface of a substrate using a vacuum system or thermal, UV, mechanical, or chemical bonding procedures. An array may also be produced by hand or by using available devices, materials, and machines, e.g. Brinkmann® multichannel pipettors or robotic instruments. The array may contain from 2 to 1,000,000 or any other feasible number of oligonucleotides.

In order to conduct sample analysis using the microarrays, polynucleotides are extracted from a sample. The sample may be obtained from any bodily fluid, e.g., blood, urine, saliva, phlegm, gastric juices, cultured cells, biopsies, or other tissue preparations. To produce probes, the polynucleotides extracted from the sample are used to produce nucleic acid sequences complementary to the nucleic acids on the microarray. If the microarray contains cDNAs, antisense RNAs (aRNAs) are appropriate probes. Therefore, in one aspect, mRNA is reverse-transcribed to cDNA. The cDNA, in the presence of fluorescent label, is used to produce fragment or oligonucleotide aRNA probes. The fluorescently labeled probes are incubated with the microarray so that the probes hybridize to the microarray oligonucleotides. Nucleic acid sequences used as probes can include polynucleotides, fragments, and complementary or antisense sequences produced using restriction enzymes, PCR, or other methods known in the art.

Hybridization conditions can be adjusted so that hybridization occurs with varying degrees of complementarity. A scanner can be used to determine the levels and patterns of fluorescence after removal of any nonhybridized probes. The degree of complementarity and the relative abundance of each oligonucleotide sequence on the microarray can be assessed through analysis of the scanned images. A detection system may be used to measure the absence, presence, or level of hybridization for any of the sequences. (See, e.g., Heller, R.A. et al. (1997) Proc. Natl. Acad. Sci. 94:2150-2155.)

In another embodiment of the invention, nucleic acid sequences encoding SIGP may be used to generate hybridization probes useful in mapping the naturally occurring genomic sequence. The sequences may be mapped to a particular chromosome, to a specific region of a chromosome, or to artificial chromosome constructions, e.g., human artificial chromosomes (HACs), yeast artificial chromosomes (YACs), bacterial artificial chromosomes (BACs), bacterial P1 constructions, or single chromosome cDNA libraries.

(See, e.g., Price, C.M. (1993) Blood Rev. 7:127-134; and Trask, B.J. (1991) Trends Genet. 7:149-154.)

Fluorescent in situ hybridization (FISH) may be correlated with other physical chromosome mapping techniques and genetic map data. (See, e.g., Heinz-Ulrich, et al. (1995) in Meyers, R.A. (ed.) Molecular Biology and Biotechnology, VCH Publishers New York, NY, pp. 965-968.) Examples of genetic map data can be found in various scientific journals or at the Online Mendelian Inheritance in Man (OMIM) site. Correlation between the location of the gene encoding SIGP on a physical chromosomal map and a specific disorder, or a predisposition to a specific disorder, may help define the region of DNA associated with that disorder. The nucleotide sequences of the invention may be used to detect differences in gene sequences among normal, carrier, and affected individuals.

In situ hybridization of chromosomal preparations and physical mapping techniques, such as linkage analysis using established chromosomal markers, may be used for extending genetic maps. Often the placement of a gene on the chromosome of another mammalian species, such as mouse, may reveal associated markers even if the number or arm of a particular human chromosome is not known. New sequences can be assigned to chromosomal arms by physical mapping. This provides valuable information to investigators searching for disease genes using positional cloning or other gene discovery techniques. Once the disease or syndrome has been crudely localized by genetic linkage to a particular genomic region, e.g., AT to 11q22-23, any sequences mapping to that area may represent associated or regulatory genes for further investigation. (See, e.g., Gatti, R.A. et al. (1988) Nature 336:577-580.) The nucleotide sequence of the subject invention may also be used to detect differences in the chromosomal location due to translocation, inversion, etc., among normal, carrier, or affected individuals.

In another embodiment of the invention, SIGP, its catalytic or immunogenic fragments, or oligopeptides thereof can be used for screening libraries of compounds in any of a variety of drug screening techniques. The fragment employed in such screening may be free in solution, affixed to a solid support, borne on a cell surface, or located intracellularly. The formation of binding complexes between SIGP and the agent being tested may be measured.

Another technique for drug screening provides for high throughput screening of compounds having suitable binding affinity to the protein of interest. (See, e.g., Geysen,

et al. (1984) PCT application WO84/03564.) In this method, large numbers of different small test compounds are synthesized on a solid substrate, such as plastic pins or some other surface. The test compounds are reacted with SIGP, or fragments thereof, and washed. Bound SIGP is then detected by methods well known in the art. Purified SIGP can also be coated directly onto plates for use in the aforementioned drug screening techniques. Alternatively, non-neutralizing antibodies can be used to capture the peptide and immobilize it on a solid support.

In another embodiment, one may use competitive drug screening assays in which neutralizing antibodies capable of binding SIGP specifically compete with a test compound for binding SIGP. In this manner, antibodies can be used to detect the presence of any peptide which shares one or more antigenic determinants with SIGP.

In additional embodiments, the nucleotide sequences which encode SIGP may be used in any molecular biology techniques that have yet to be developed, provided the new techniques rely on properties of nucleotide sequences that are currently known, including, but not limited to, such properties as the triplet genetic code and specific base pair interactions.

The examples below are provided to illustrate the subject invention and are not included for the purpose of limiting the invention.

EXAMPLES

For purposes of example, the preparation and sequencing of the SPLNNOT04 cDNA library, from which Incyte Clones 1534876 and 1559131 were isolated, is described. Preparation and sequencing of cDNAs in libraries in the LIFESEQ™ database have varied over time, and the gradual changes involved use of kits, plasmids, and machinery available at the particular time the library was made and analyzed.

I. SPLNNOT04 cDNA Library Construction

The SPLNNOT04 cDNA library was constructed from microscopically normal spleen tissue obtained from a 2-year-old Hispanic male who died of cerebral anoxia. The patient's serologies and past medical history were negative.

The frozen tissue was homogenized and lysed using a Brinkmann Homogenizer Polytron PT-3000 (Brinkmann Instruments, Westbury, NJ) in guanidinium isothiocyanate

solution. The lysate was centrifuged over a 5.7 M CsCl cushion using an Beckman SW28 rotor in a Beckman L8-70M Ultracentrifuge (Beckman Instruments) for 18 hours at 25,000 rpm at ambient temperature. The RNA was extracted with acid phenol pH 4.0, precipitated using 0.3 M sodium acetate and 2.5 volumes of ethanol, resuspended in RNAse-free water and DNase treated at 37°C. The RNA extraction and precipitation were repeated as before. The mRNA was then isolated using the Qiagen Oligotex kit (QIAGEN Inc., Chatsworth, CA) and used to construct the cDNA library.

The mRNA was handled according to the recommended protocols in the SuperScript plasmid system (Cat. #18248-013, GIBCO-BRL, Gaithersburg, MD). cDNA synthesis was initiated with a NotI-oligo d(T) primer. Double-stranded cDNA was blunted, ligated to EcoRI adaptors, digested with NotI, fractionated on a Sepharose CL4B column (Cat. #275105-01, Pharmacia), and those cDNAs exceeding 400 bp were ligated into the NotI and EcoRI sites of the pINCY 1 vector (Incyte). The plasmid pINCY 1 was subsequently transformed into DH5α™ competent cells (Cat. #18258-012, GIBCO-BRL).

II Isolation and Sequencing of cDNA Clones

Plasmid cDNA was released from the cells and purified using the REAL Prep 96 plasmid kit (Catalog #26173, QIAGEN). The recommended protocol was employed except for the following changes: 1) the bacteria were cultured in 1 ml of sterile Terrific Broth (Catalog #22711, GIBCO-BRL) with carbenicillin at 25 mg/L and glycerol at 0.4%; 2) after inoculation, the cultures were incubated for 19 hours and at the end of incubation, the cells were lysed with 0.3 ml of lysis buffer; and 3) following isopropanol precipitation, the plasmid DNA pellet was resuspended in 0.1 ml of distilled water. After the last step in the protocol, samples were transferred to a 96-well block for storage at 4° C.

cDNAs were sequenced according to the method of Sanger et al. (1975, J. Mol. Biol. 94:441f), using the Perkin Elmer Catalyst 800 or a Hamilton Micro Lab 2200 (Hamilton, Reno, NV) in combination with Peltier Thermal Cyclers (PTC200 from MJ Research, Watertown, MA) and Applied Biosystems 377 DNA Sequencing Systems or the Perkin Elmer 373 DNA Sequencing System and the reading frame was determined.

III. Homology Searching of cDNA Clones and Their Deduced Proteins

The nucleotide sequences and/or amino acid sequences of the Sequence Listing were

used to query sequences in the GenBank, SwissProt, BLOCKS, and Pima II databases. These databases, which contain previously identified and annotated sequences, were searched for regions of homology using BLAST (Basic Local Alignment Search Tool). (See, e.g., Altschul, S.F. (1993) J. Mol. Evol 36:290-300; and Altschul et al. (1990) J. Mol. Biol. 215:403-410.)

BLAST produced alignments of both nucleotide and amino acid sequences to determine sequence similarity. Because of the local nature of the alignments, BLAST was especially useful in determining exact matches or in identifying homologs which may be of prokaryotic (bacterial) or eukaryotic (animal, fungal, or plant) origin. Other algorithms could have been used when dealing with primary sequence patterns and secondary structure gap penalties. (See, e.g., Smith, T. et al. (1992) Protein Engineering 5:35-51.) The sequences disclosed in this application have lengths of at least 49 nucleotides and have no more than 12% uncalled bases (where N is recorded rather than A, C, G, or T).

The BLAST approach searched for matches between a query sequence and a database sequence. BLAST evaluated the statistical significance of any matches found, and reported only those matches that satisfy the user-selected threshold of significance. In this application, threshold was set at 10^{-25} for nucleotides and 10^{-8} for peptides.

Incyte nucleotide sequences were searched against the GenBank databases for primate (pri), rodent (rod), and other mammalian sequences (mam), and deduced amino acid sequences from the same clones were then searched against GenBank functional protein databases, mammalian (mamp), vertebrate (vrtp), and eukaryote (eukp), for homology.

IV. Northern Analysis

Northern analysis is a laboratory technique used to detect the presence of a transcript of a gene and involves the hybridization of a labeled nucleotide sequence to a membrane on which RNAs from a particular cell type or tissue have been bound. (See, e.g., Sambrook, supra, ch. 7; and Ausubel, F.M. et al. supra, ch. 4 and 16.)

Analogous computer techniques applying BLAST are used to search for identical or related molecules in nucleotide databases such as GenBank or LIFESEQ™ database (Incyte Pharmaceuticals). This analysis is much faster than multiple membrane-based hybridizations. In addition, the sensitivity of the computer search can be modified to determine whether any particular match is categorized as exact or homologous.

The basis of the search is the product score, which is defined as:

$$\frac{\%\text{ sequence identity} \times \%\text{ maximum BLAST score}}{100}$$

The product score takes into account both the degree of similarity between two sequences and the length of the sequence match. For example, with a product score of 40, the match will be exact within a 1% to 2% error, and, with a product score of 70, the match will be exact. Homologous molecules are usually identified by selecting those which show product scores between 15 and 40, although lower scores may identify related molecules.

The results of northern analysis are reported as a list of libraries in which the transcript encoding SIGP occurs. Abundance and percent abundance are also reported. Abundance directly reflects the number of times a particular transcript is represented in a cDNA library, and percent abundance is abundance divided by the total number of sequences examined in the cDNA library.

V. Extension of SIGP Encoding Polynucleotides

The nucleic acid sequence of one of the polynucleotides of the present invention was used to design oligonucleotide primers for extending a partial nucleotide sequence to full length. One primer was synthesized to initiate extension of an antisense polynucleotide, and the other was synthesized to initiate extension of a sense polynucleotide. Primers were used to facilitate the extension of the known sequence "outward" generating amplicons containing new unknown nucleotide sequence for the region of interest. The initial primers were designed from the cDNA using OLIGO 4.06 (National Biosciences, Plymouth, MN), or another appropriate program, to be about 22 to 30 nucleotides in length, to have a GC content of about 50% or more, and to anneal to the target sequence at temperatures of about 68°C to about 72°C. Any stretch of nucleotides which would result in hairpin structures and primer-primer dimerizations was avoided.

Selected human cDNA libraries (GIBCO/BRL) were used to extend the sequence. If more than one extension is necessary or desired, additional sets of primers are designed to further extend the known region.

High fidelity amplification was obtained by following the instructions for the XL-PCR kit (Perkin Elmer) and thoroughly mixing the enzyme and reaction mix. PCR was performed using the Peltier Thermal Cycler (PTC200; M.J. Research, Watertown, MA),

beginning with 40 pmol of each primer and the recommended concentrations of all other components of the kit, with the following parameters:

Step 1	94° C for 1 min (initial denaturation)
Step 2	65° C for 1 min
Step 3	68° C for 6 min
Step 4	94° C for 15 sec
Step 5	65° C for 1 min
Step 6	68° C for 7 min
Step 7	Repeat steps 4 through 6 for an additional 15 cycles
Step 8	94° C for 15 sec
Step 9	65° C for 1 min
Step 10	68° C for 7:15 min
Step 11	Repeat steps 8 through 10 for an additional 12 cycles
Step 12	72° C for 8 min
Step 13	4° C (and holding)

A 5 μl to 10 μl aliquot of the reaction mixture was analyzed by electrophoresis on a low concentration (about 0.6% to 0.8%) agarose mini-gel to determine which reactions were successful in extending the sequence. Bands thought to contain the largest products were excised from the gel, purified using QIAQuick™ (QIAGEN Inc., Chatsworth, CA), and trimmed of overhangs using Klenow enzyme to facilitate religation and cloning.

After ethanol precipitation, the products were redissolved in 13 μl of ligation buffer, 1μl T4-DNA ligase (15 units) and 1μl T4 polynucleotide kinase were added, and the mixture was incubated at room temperature for 2 to 3 hours, or overnight at 16° C. Competent E. coli cells (in 40 μl of appropriate media) were transformed with 3 μl of ligation mixture and cultured in 80 μl of SOC medium. (See, e.g., Sambrook, supra, Appendix A, p. 2.) After incubation for one hour at 37° C, the E. coli mixture was plated on Luria Bertani (LB) agar (See, e.g., Sambrook, supra, Appendix A, p. 1) containing 2x Carb. The following day, several colonies were randomly picked from each plate and cultured in 150 μl of liquid LB/2x Carb medium placed in an individual well of an appropriate commercially-available sterile 96-well microtiter plate. The following day, 5 μl of each overnight culture was transferred into a non-sterile 96-well plate and, after dilution 1:10 with water, 5 μl from each sample was transferred into a PCR array.

For PCR amplification, 18 μl of concentrated PCR reaction mix (3.3x) containing 4 units of rTth DNA polymerase, a vector primer, and one or both of the gene specific primers used for the extension reaction were added to each well. Amplification was

performed using the following conditions:

Step 1	94° C for 60 sec
Step 2	94° C for 20 sec
Step 3	55° C for 30 sec
Step 4	72° C for 90 sec
Step 5	Repeat steps 2 through 4 for an additional 29 cycles
Step 6	72° C for 180 sec
Step 7	4° C (and holding)

Aliquots of the PCR reactions were run on agarose gels together with molecular weight markers. The sizes of the PCR products were compared to the original partial cDNAs, and appropriate clones were selected, ligated into plasmid, and sequenced.

In like manner, the nucleotide sequence of one of the nucleotide sequences of the present invention were used to obtain 5' regulatory sequences using the procedure above, oligonucleotides designed for 5' extension, and an appropriate genomic library.

VI. Labeling and Use of Individual Hybridization Probes

Hybridization probes derived from one of the nucleotide sequences of the present invention are employed to screen cDNAs, genomic DNAs, or mRNAs. Although the labeling of oligonucleotides, consisting of about 20 base pairs, is specifically described, essentially the same procedure is used with larger nucleotide fragments. Oligonucleotides are designed using state-of-the-art software such as OLIGO 4.06 (National Biosciences) and labeled by combining 50 pmol of each oligomer, 250 μCi of [γ-^{32}P] adenosine triphosphate (Amersham, Chicago, IL), and T4 polynucleotide kinase (DuPont NEN®, Boston, MA). The labeled oligonucleotides are substantially purified using a Sephadex G-25 superfine resin column (Pharmacia & Upjohn, Kalamazoo, MI). An aliquot containing 10^7 counts per minute of the labeled probe is used in a typical membrane-based hybridization analysis of human genomic DNA digested with one of the following endonucleases: Ase I, Bgl II, Eco RI, Pst I, Xba 1, or Pvu II (DuPont NEN, Boston, MA).

The DNA from each digest is fractionated on a 0.7 percent agarose gel and transferred to nylon membranes (Nytran Plus, Schleicher & Schuell, Durham, NH). Hybridization is carried out for 16 hours at 40°C. To remove nonspecific signals, blots are sequentially washed at room temperature under increasingly stringent conditions up to 0.1 x saline sodium citrate and 0.5% sodium dodecyl sulfate. After XOMAT AR™ film

(Kodak, Rochester, NY) is exposed to the blots to film for several hours, hybridization patterns are compared visually.

VII. Microarrays

To produce oligonucleotides for a microarray, one of the nucleotide sequences of the present invention is examined using a computer algorithm which starts at the 3' end of the nucleotide sequence. For each, the algorithm identifies oligomers of defined length that are unique to the nucleic acid sequence, have a GC content within a range suitable for hybridization, and lack secondary structure that would interfere with hybridization. The algorithm identifies approximately 20 oligonucleotides corresponding to each nucleic acid sequence. For each sequence-specific oligonucleotide, a pair of oligonucleotides is synthesized in which the first oligonucleotides differs from the second oligonucleotide by one nucleotide in the center of the sequence. The oligonucleotide pairs can be arranged on a substrate, e.g. a silicon chip, using a light-directed chemical process. (See, e.g., Chee, supra.)

In the alternative, a chemical coupling procedure and an ink jet device can be used to synthesize oligomers on the surface of a substrate. (See, e.g., Baldeschweiler, supra.) An array analogous to a dot or slot blot may also be used to arrange and link fragments or oligonucleotides to the surface of a substrate using or thermal, UV, mechanical, or chemical bonding procedures, or a vacuum system. A typical array may be produced by hand or using available methods and machines and contain any appropriate number of elements. After hybridization, nonhybridized probes are removed and a scanner used to determine the levels and patterns of fluorescence. The degree of complementarity and the relative abundance of each oligonucleotide sequence on the microarray may be assessed through analysis of the scanned images.

VIII. Complementary Polynucleotides

Sequences complementary to the SIGP-encoding sequences, or any parts thereof, are used to detect, decrease, or inhibit expression of naturally occurring SIGP. Although use of oligonucleotides comprising from about 15 to 30 base pairs is described, essentially the same procedure is used with smaller or with larger sequence fragments. Appropriate oligonucleotides are designed using Oligo 4.06 software and the coding sequence of SIGP.

To inhibit transcription, a complementary oligonucleotide is designed from the most unique 5' sequence and used to prevent promoter binding to the coding sequence. To inhibit translation, a complementary oligonucleotide is designed to prevent ribosomal binding to the SIGP-encoding transcript.

IX. Expression of SIGP

Expression of SIGP is accomplished by subcloning the cDNA into an appropriate vector and transforming the vector into host cells. This vector contains an appropriate promoter, e.g., ß-galactosidase upstream of the cloning site, operably associated with the cDNA of interest. (See, e.g.,Sambrook, supra, pp. 404-433; and Rosenberg, M. et al. (1983) Methods Enzymol. 101:123-138.)

Induction of an isolated, transformed bacterial strain with isopropyl beta-D-thiogalactopyranoside (IPTG) using standard methods produces a fusion protein which consists of the first 8 residues of ß-galactosidase, about 5 to 15 residues of linker, and the full length protein. The signal residues direct the secretion of SIGP into bacterial growth media which can be used directly in the following assay for activity.

X. Production of SIGP Specific Antibodies

SIGP substantially purified using PAGE electrophoresis (see, e.g., Harrington, M.G. (1990) Methods Enzymol. 182:488-495), or other purification techniques, is used to immunize rabbits and to produce antibodies using standard protocols. The SIGP amino acid sequence is analyzed using DNASTAR software (DNASTAR Inc) to determine regions of high immunogenicity, and a corresponding oligopeptide is synthesized and used to raise antibodies by means known to those of skill in the art. Methods for selection of appropriate epitopes, such as those near the C-terminus or in hydrophilic regions are well described in the art. (See, e.g., Ausubel et al. supra, ch. 11.)

Typically, the oligopeptides are 15 residues in length, and are synthesized using an Applied Biosystems Peptide Synthesizer Model 431A using fmoc-chemistry and coupled to KLH (Sigma, St. Louis, MO) by reaction with N-maleimidobenzoyl-N-hydroxysuccinimide ester (MBS) to increase immunogenicity. (See, e.g., Ausubel et al. supra.) Rabbits are immunized with the oligopeptide-KLH complex in complete Freund's adjuvant. Resulting antisera are tested for antipeptide activity, for example, by binding the

peptide to plastic, blocking with 1% BSA, reacting with rabbit antisera, washing, and reacting with radio-iodinated goat anti-rabbit IgG.

XI. Purification of Naturally Occurring SIGP Using Specific Antibodies

Naturally occurring or recombinant SIGP is substantially purified by immunoaffinity chromatography using antibodies specific for SIGP. An immunoaffinity column is constructed by covalently coupling anti-SIGP antibody to an activated chromatographic resin, such as CNBr-activated Sepharose (Pharmacia & Upjohn). After the coupling, the resin is blocked and washed according to the manufacturer's instructions.

Media containing SIGP are passed over the immunoaffinity column, and the column is washed under conditions that allow the preferential absorbance of SIGP (e.g., high ionic strength buffers in the presence of detergent). The column is eluted under conditions that disrupt antibody/SIGP binding (e.g., a buffer of pH 2 to pH 3, or a high concentration of a chaotrope, such as urea or thiocyanate ion), and SIGP is collected.

XII. Identification of Molecules Which Interact with SIGP

SIGP, or biologically active fragments thereof, are labeled with ^{125}I Bolton-Hunter reagent. (See, e.g., Bolton et al. (1973) Biochem. J. 133:529.) Candidate molecules previously arrayed in the wells of a multi-well plate are incubated with the labeled SIGP, washed, and any wells with labeled SIGP complex are assayed. Data obtained using different concentrations of SIGP are used to calculate values for the number, affinity, and association of SIGP with the candidate molecules.

Various modifications and variations of the described methods and systems of the invention will be apparent to those skilled in the art without departing from the scope and spirit of the invention. Although the invention has been described in connection with specific preferred embodiments, it should be understood that the invention as claimed should not be unduly limited to such specific embodiments. Indeed, various modifications of the described modes for carrying out the invention which are obvious to those skilled in molecular biology or related fields are intended to be within the scope of the following claims.

What is claimed is:

1. A substantially purified human signal peptide-containing protein (SIGP) comprising a polypeptide having an amino acid sequence selected from the group consisting of SEQ ID NO:1, SEQ ID NO:2, SEQ ID NO:3, SEQ ID NO:4, SEQ ID NO:5, SEQ ID NO:6, SEQ ID NO:7, SEQ ID NO:8, SEQ ID NO:9, SEQ ID NO:10, SEQ ID NO:11, SEQ ID NO:12, SEQ ID NO:13, SEQ ID NO:14, and SEQ ID NO:15.

2. An isolated and purified polynucleotide which hybridizes under stringent conditions to the polynucleotide encoding an SIGP of claim 1.

3. An isolated and purified polynucleotide encoding the SIGP of claim 1.

4. A microarray containing at least a fragment of at least one of the polynucleotides encoding an SIGP of claim 1.

5. An isolated and purified polynucleotide variant having at least 90% polynucleotide identity to the polynucleotide of claim 3.

6. A composition comprising the polynucleotide of claim 3.

7. An isolated and purified polynucleotide which hybridizes under stringent conditions to the polynucleotide of claim 3.

8. An isolated and purified polynucleotide which is complementary to the polynucleotide of claim 3.

9. An isolated and purified polynucleotide having a nucleic acid sequence selected from the group consisting of SEQ ID NO:16, SEQ ID NO:17, SEQ ID NO:18, SEQ ID NO:19, SEQ ID NO:20, SEQ ID NO:21, SEQ ID NO:22, SEQ ID NO:23, SEQ ID NO:24, SEQ ID NO:25, SEQ ID NO:26, SEQ ID NO:27, SEQ ID NO:28, SEQ ID NO:29, and SEQ ID NO:30.

10. An isolated and purified polynucleotide variant having at least 90% polynucleotide identity to the polynucleotide of claim 9.

11. An isolated and purified polynucleotide which is complementary to the polynucleotide sequence of claim 9.

12. An expression vector containing at least a fragment of the polynucleotide of claim 3.

13. A host cell containing the expression vector of claim 12.

14. A method for producing a polypeptide encoding a human signal peptide-containing protein, the method comprising the steps of:

(a) culturing the host cell of claim 13 under conditions suitable for the expression of the polypeptide; and

(b) recovering the polypeptide from the host cell culture.

15. A pharmaceutical composition comprising the SIGP of claim 1 in conjunction with a suitable pharmaceutical carrier.

16. A purified antibody which specifically binds to the SIGP of claim 1.

17. A purified agonist of the SIGP of claim 1.

18. A purified antagonist of the SIGP of claim 1.

19. A method for treating or preventing a cancer, the method comprising administering to a subject in need of such treatment an effective amount of the pharmaceutical composition of claim 15.

20. A method for treating or preventing a cancer, the method comprising administering to a subject in need of such treatment an effective amount of the antagonist of claim 18.

21. A method for treating or preventing an immune response, the method comprising administering to a subject in need of such treatment an effective amount of the antagonist of claim 18.

22. A method for detecting a polynucleotide encoding a human signal peptide-containing protein in a biological sample containing nucleic acids, the method comprising the steps of:

(a) hybridizing the polynucleotide of claim 8 to at least one of the nucleic acids of the biological sample, thereby forming a hybridization complex; and

(b) detecting the hybridization complex, wherein the presence of the hybridization complex correlates with the presence of a polynucleotide encoding SIGP in the biological sample.

23. The method of claim 22 wherein the nucleic acids of the biological sample are amplified by the polymerase chain reaction prior to the hybridizing step.

SEQUENCE LISTING

<110> INCYTE PHARMACEUTICALS, INC.
LAL, Preeti
HILLMAN, Jennifer L.
CORLEY, Neil C.
GUEGLER, Karl J.
BAUGHN, Mariah R.
SATHER, Susan K.
SHAH, Purvi

<120> HUMAN SIGNAL PEPTIDE-CONTAINING PROTEINS

<130> PF-0459 PCT

<140> To Be Assigned
<141> Herewith

<150> 09/002,485
<151> 1997-12-31

<160> 30

<170> PERL PROGRAM

<210> 1
<211> 236
<212> PRT
<213> Homo sapiens

<220> -
<223> 866885

<400> 1

```
Met Ala Pro Asp Pro Trp Phe Ser Thr Tyr Asp Ser Thr Cys Gln
  1               5                  10                  15
Ile Ala Gln Glu Ile Ala Glu Lys Ile Gln Gln Arg Asn Gln Tyr
             20                  25                  30
Glu Arg Lys Gly Glu Lys Ala Pro Lys Leu Thr Val Thr Ile Arg
             35                  40                  45
Ala Leu Leu Gln Asn Leu Lys Glu Lys Ile Ala Leu Leu Lys Asp
             50                  55                  60
Leu Leu Leu Arg Ala Val Ser Thr His Gln Ile Thr Gln Leu Glu
             65                  70                  75
Gly Asp Arg Arg Gln Asn Leu Leu Asp Asp Leu Val Thr Arg Glu
             80                  85                  90
Arg Leu Leu Leu Ala Ser Phe Lys Asn Glu Gly Ala Glu Pro Asp
             95                 100                 105
Leu Ile Arg Ser Ser Leu Met Ser Glu Glu Ala Lys Arg Gly Ala
            110                 115                 120
Pro Asn Pro Trp Leu Phe Glu Glu Pro Glu Glu Thr Arg Gly Leu
            125                 130                 135
Gly Phe Asp Glu Ile Arg Gln Gln Gln Gln Gln Lys Ile Ile Gln Glu
            140                 145                 150
```

```
Gln Asp Ala Gly Leu Asp Ala Leu Ser Ser Ile Ile Ser Arg Gln
                155                 160                 165
Lys Gln Met Gly Gln Glu Ile Gly Asn Glu Leu Asp Glu Gln Asn
                170                 175                 180
Glu Ile Ile Asp Asp Leu Ala Asn Leu Val Glu Asn Thr Asp Glu
                185                 190                 195
Lys Leu Arg Asn Glu Thr Arg Arg Val Asn Met Val Asp Arg Lys
                200                 205                 210
Ser Ala Ser Cys Gly Met Ile Met Val Ile Leu Leu Leu Leu Val
                215                 220                 225
Ala Ile Val Val Val Ala Val Trp Pro Thr Asn
                230                 235
```

<210> 2
<211> 267
<212> PRT
<213> Homo sapiens

<220> -
<223> 1273453

<400> 2

```
Met Val Ile Ser Trp His Leu Ala Ser Asp Met Asp Cys Val Val
  1               5                  10                  15
Thr Leu Thr Thr Asp Ala Ala Arg Arg Ile Tyr Asp Glu Thr Gln
                20                  25                  30
Gly Arg Gln Gln Val Leu Pro Leu Asp Ser Ile Tyr Lys Lys Thr
                35                  40                  45
Leu Pro Asp Trp Lys Arg Ser Leu Pro His Phe Arg Asn Gly Lys
                50                  55                  60
Leu Tyr Phe Lys Pro Ile Gly Asp Pro Val Phe Ala Arg Asp Leu
                65                  70                  75
Leu Thr Phe Pro Asp Asn Val Glu His Cys Glu Thr Val Phe Gly
                80                  85                  90
Met Leu Leu Gly Asp Thr Ile Ile Leu Asp Asn Leu Asp Ala Ala
                95                 100                 105
Asn His Tyr Arg Lys Glu Val Val Lys Ile Thr His Cys Pro Thr
                110                 115                 120
Leu Leu Thr Arg Asp Gly Asp Arg Ile Arg Ser Asn Gly Lys Phe
                125                 130                 135
Gly Gly Leu Gln Asn Lys Ala Pro Pro Met Asp Lys Leu Arg Gly
                140                 145                 150
Met Val Phe Gly Ala Pro Val Pro Lys Gln Cys Leu Ile Leu Gly
                155                 160                 165
Glu Gln Ile Asp Leu Leu Gln Gln Tyr Arg Ser Ala Val Cys Lys
                170                 175                 180
Leu Asp Ser Val Asn Lys Asp Leu Asn Ser Gln Leu Glu Tyr Leu
                185                 190                 195
Arg Thr Pro Asp Met Arg Lys Lys Lys Gln Glu Leu Asp Glu His
                200                 205                 210
Glu Lys Asn Leu Lys Leu Ile Glu Glu Lys Leu Gly Met Thr Pro
                215                 220                 225
Ile Arg Lys Cys Asn Asp Ser Leu Arg His Ser Pro Lys Val Glu
                230                 235                 240
Thr Thr Asp Cys Pro Val Pro Pro Lys Arg Met Arg Arg Glu Ala
```

```
                245                 250                 255
Thr Arg Gln Asn Arg Ile Ile Thr Lys Thr Asp Val
                260                 265
```

<210> 3
<211> 161
<212> PRT
<213> Homo sapiens

<220> -
<223> 1534876

<400> 3

```
Met Trp Phe Leu Gly Cys Thr Gly Pro Gly Cys Gly Cys Ala Gly
  1               5                  10                  15
Val Cys Lys Val Val Pro Cys Ile Ser Thr Gly Phe Glu Thr Ser
                20                  25                  30
Gly Pro Cys Pro Ser Ser Arg Glu Gly Phe Leu Phe Phe Leu Thr
                35                  40                  45
Gln Val Thr Phe Gln Pro Phe Gln Phe Pro Ser Phe Ser Ala Leu
                50                  55                  60
Pro Ser Asn Ser Ala Asn Pro Gly Val Gly Ser Gln Gly Gly Arg
                65                  70                  75
Glu Cys Pro Thr Thr Phe Ser Gly Gln Pro Leu Thr Pro Lys Pro
                80                  85                  90
Leu Pro Pro Ser Ile Leu His Pro Leu Pro Ile Gln Pro Lys Cys
                95                 100                 105
Pro Gln Leu Gly Leu Ser Cys Ile Pro Val Glu Gly Pro Leu Pro
               110                 115                 120
Cys Leu Ser Glu Val Arg Leu Cys Cys Val Met Gly Arg Leu Cys
               125                 130                 135
Pro Ser Pro Pro Leu Ala Arg Cys Thr Cys Phe Leu Val Cys Thr
               140                 145                 150
Arg Cys Pro Gly Gly Pro Ser Leu Pro Cys Gln
               155                 160
```

<210> 4
<211> 150
<212> PRT
<213> Homo sapiens

<220> -
<223> 1634813

<400> 4

```
Met Asn Leu Trp Leu Leu Ala Cys Leu Val Ala Gly Phe Leu Gly
  1               5                  10                  15
Ala Trp Ala Pro Ala Val His Ala Gln Gly Val Phe Glu Asp Cys
                20                  25                  30
Cys Leu Ala Tyr His Tyr Pro Ile Gly Trp Ala Val Leu Arg Arg
                35                  40                  45
```

```
Ala Trp Thr Tyr Arg Ile Gln Glu Val Ser Gly Ser Cys Asn Leu
                50                  55                  60
Pro Ala Ala Ile Phe Tyr Leu Pro Lys Arg His Arg Lys Val Cys
                65                  70                  75
Gly Asn Pro Lys Ser Arg Glu Val Gln Arg Ala Met Lys Leu Leu
                80                  85                  90
Asp Ala Arg Asn Lys Val Phe Ala Lys Leu Arg His Asn Thr Gln
                95                 100                 105
Thr Phe Gln Ala Gly Pro His Ala Val Lys Lys Leu Ser Ser Gly
               110                 115                 120
Asn Ser Lys Leu Ser Ser Ser Lys Phe Ser Asn Pro Ile Ser Ser
               125                 130                 135
Ser Lys Arg Asn Val Ser Leu Leu Ile Ser Ala Asn Ser Gly Leu
               140                 145                 150
```

<210> 5
<211> 118
<212> PRT
<213> Homo sapiens

<220> -
<223> 1711840

<400> 5

```
Met Gln His Arg Gly Phe Leu Leu Leu Thr Leu Leu Ala Leu Leu
  1                 5                  10                  15
Ala Leu Thr Ser Ala Val Ala Lys Lys Gln Asp Lys Val Lys Lys
                20                  25                  30
Gly Gly Pro Gly Ser Glu Cys Ala Glu Trp Ala Trp Gly Pro Cys
                35                  40                  45
Thr Pro Ser Ser Lys Gly Phe Ala Ala Val Gly Phe Pro Arg Gly
                50                  55                  60
Pro Pro Trp Gly Gly Pro Arg Thr Gln Pro Ala Val Leu Val Glu
                65                  70                  75
Arg Val Ala Pro Gly Lys Leu Glu Arg Lys Glu Phe Trp Ala Pro
                80                  85                  90
Gly Leu Trp Lys Val Gly Gln Ile Phe Trp Lys Lys Thr Trp Arg
                95                 100                 105
Val Cys Arg Ser Val Lys Trp Gly Arg Gly Gln Lys Asn
               110                 115
```

<210> 6
<211> 248
<212> PRT
<213> Homo sapiens

<220> -
<223> 1747327

<400> 6

```
Met Gln Thr Cys Pro Leu Ala Phe Pro Gly His Val Ser Gln Ala
  1                 5                  10                  15
Leu Gly Thr Leu Leu Phe Leu Ala Ala Ser Leu Ser Ala Gln Asn
                20                  25                  30
```

```
Glu Gly Trp Asp Ser Pro Ile Cys Thr Glu Gly Val Val Ser Val
                35                  40                  45
Ser Trp Gly Glu Asn Thr Val Met Ser Cys Asn Ile Ser Asn Ala
                50                  55                  60
Phe Ser His Val Asn Ile Lys Leu Arg Ala His Gly Gln Glu Ser
                65                  70                  75
Ala Ile Phe Asn Glu Val Ala Pro Gly Tyr Phe Ser Arg Asp Gly
                80                  85                  90


Trp Gln Leu Gln Val Gln Gly Gly Val Ala Gln Leu Val Ile Lys
                95                 100                 105
Gly Ala Arg Asp Ser His Ala Gly Leu Tyr Met Trp His Leu Val
               110                 115                 120
Gly His Gln Arg Asn Asn Arg Gln Val Thr Leu Glu Val Ser Gly
               125                 130                 135
Ala Glu Pro Gln Ser Ala Pro Asp Thr Gly Phe Trp Pro Val Pro
               140                 145                 150
Ala Val Val Thr Ala Val Phe Ile Leu Leu Val Ala Leu Val Met
               155                 160                 165
Phe Ala Trp Tyr Arg Cys Arg Cys Ser Gln Gln Arg Arg Glu Lys
               170                 175                 180
Lys Phe Phe Leu Leu Glu Pro Gln Met Lys Val Ala Ala Leu Arg
               185                 190                 195
Ala Gly Ala Gln Gln Gly Leu Ser Arg Ala Ser Ala Glu Leu Trp
               200                 205                 210
Thr Pro Asp Ser Glu Pro Thr Pro Arg Pro Leu Ala Leu Val Phe
               215                 220                 225
Lys Pro Ser Pro Leu Gly Ala Leu Glu Leu Leu Ser Pro Gln Pro
               230                 235                 240
Leu Phe Pro Tyr Ala Ala Asp Pro
               245


<210> 7
<211> 404
<212> PRT
<213> Homo sapiens

<220> -
<223> 1864292

<400> 7
Met Lys Met Glu Glu Ala Val Gly Lys Val Glu Glu Leu Ile Glu
  1                 5                  10                  15
Ser Glu Ala Pro Pro Lys Ala Ser Glu Gln Glu Thr Ala Lys Glu
                20                  25                  30
Glu Asp Gly Ser Val Glu Leu Glu Ser Gln Val Gln Lys Asp Gly
                35                  40                  45
Val Ala Asp Ser Thr Val Ile Ser Ser Met Pro Cys Leu Leu Met
                50                  55                  60
Glu Leu Arg Arg Asp Ser Ser Glu Ser Gln Leu Ala Ser Thr Glu
                65                  70                  75
Ser Asp Lys Pro Thr Thr Gly Arg Val Tyr Glu Ser Asp Pro Ser
                80                  85                  90
Asn His Cys Met Leu Ser Pro Ser Ser Ser Gly His Leu Ala Asp
```

```
                 95                  100                 105
Ser Asp Thr Leu Ser Ser Ala Glu Glu Asn Glu Pro Ser Gln Ala
                110                 115                 120
Glu Thr Ala Val Glu Gly Asp Pro Ser Gly Val Ser Gly Ala Thr
                125                 130                 135
Val Gly Arg Lys Ser Arg Arg Ser Arg Ser Glu Ser Glu Thr Ser
                140                 145                 150
Thr Met Ala Ala Lys Lys Asn Arg Gln Ser Ser Asp Lys Gln Asn
                155                 160                 165
Gly Arg Val Ala Lys Val Lys Gly His Arg Ser Gln Lys His Lys
                170                 175                 180

Glu Arg Ile Arg Leu Leu Arg Gln Lys Arg Glu Ala Ala Ala Arg
                185                 190                 195
Lys Lys Tyr Asn Leu Leu Gln Asp Ser Ser Thr Ser Asp Ser Asp
                200                 205                 210
Leu Thr Cys Asp Ser Ser Thr Ser Ser Ser Asp Asp Asp Glu Glu
                215                 220                 225
Val Ser Gly Ser Ser Lys Thr Ile Thr Ala Glu Ile Pro Asp Gly
                230                 235                 240
Pro Pro Val Val Ala His Tyr Asp Met Ser Asp Thr Asn Ser Asp
                245                 250                 255
Pro Glu Val Val Asn Val Asp Asn Leu Leu Ala Ala Ala Val Val
                260                 265                 270
Gln Glu His Ser Asn Ser Val Gly Gly Gln Asp Thr Gly Ala Thr
                275                 280                 285
Trp Arg Thr Ser Gly Leu Leu Glu Glu Leu Asn Ala Glu Ala Gly
                290                 295                 300
His Leu Asp Pro Gly Phe Leu Ala Ser Asp Lys Thr Ser Ala Gly
                305                 310                 315
Asn Ala Pro Leu Asn Glu Glu Ile Asn Ile Ala Ser Ser Asp Ser
                320                 325                 330
Glu Val Glu Ile Val Gly Val Gln Glu His Ala Arg Cys Val His
                335                 340                 345
Pro Arg Gly Gly Val Ile Gln Ser Val Ser Ser Trp Lys His Gly
                350                 355                 360
Ser Gly Thr Gln Tyr Val Ser Thr Arg Gln Thr Gln Ser Trp Thr
                365                 370                 375
Ala Val Thr Pro Gln Gln Thr Trp Ala Ser Pro Ala Glu Val Val
                380                 385                 390
Asp Leu Thr Leu Asp Glu Asp Ser Arg Arg Lys Tyr Leu Leu
                395                 400


<210> 8
<211> 405
<212> PRT
<213> Homo sapiens

<220> -
<223> 1866437

<400> 8
Met Phe Val Gln Glu Glu Lys Ile Phe Ala Gly Lys Val Leu Arg
  1               5                  10                  15
Leu His Ile Cys Ala Ser Asp Gly Ala Glu Trp Leu Glu Glu Ala
```

```
                    20                  25                  30
Thr Glu Asp Thr Ser Val Glu Lys Leu Lys Glu Arg Cys Leu Lys
                    35                  40                  45
His Cys Ala His Gly Ser Leu Glu Asp Pro Lys Ser Ile Thr His
                    50                  55                  60
His Lys Leu Ile His Ala Ala Ser Glu Arg Val Leu Ser Asp Ala
                    65                  70                  75
Arg Thr Ile Leu Glu Glu Asn Ile Gln Asp Gln Asp Val Leu Leu
                    80                  85                  90
Leu Lys Lys Lys Arg Ala Pro Ser Pro Leu Pro Lys Met Ala Asp
                    95                  100                 105
Val Ser Ala Glu Glu Lys Lys Lys Gln Asp Gln Lys Ala Pro Asp
                    110                 115                 120

Lys Glu Ala Ile Leu Arg Ala Thr Ala Asn Leu Pro Ser Tyr Asn
                    125                 130                 135
Met Asp Arg Ala Ala Val Gln Thr Asn Met Arg Asp Phe Gln Thr
                    140                 145                 150
Glu Leu Arg Lys Ile Leu Val Ser Leu Ile Glu Val Ala Gln Lys
                    155                 160                 165
Leu Leu Ala Leu Asn Pro Asp Ala Val Glu Leu Phe Lys Lys Ala
                    170                 175                 180
Asn Ala Met Leu Asp Glu Asp Glu Asp Glu Arg Val Asp Glu Ala
                    185                 190                 195
Ala Leu Arg Gln Leu Thr Glu Met Gly Phe Pro Glu Asn Arg Ala
                    200                 205                 210
Thr Lys Ala Leu Gln Leu Asn His Met Ser Val Pro Gln Ala Met
                    215                 220                 225
Glu Trp Leu Ile Glu His Ala Glu Asp Pro Thr Ile Asp Thr Pro
                    230                 235                 240
Leu Pro Gly Gln Ala Pro Pro Glu Ala Glu Gly Ala Thr Ala Ala
                    245                 250                 255
Ala Ser Glu Ala Ala Ala Gly Ala Ser Ala Thr Asp Glu Glu Ala
                    260                 265                 270
Arg Asp Glu Leu Thr Glu Ile Phe Lys Lys Ile Arg Arg Lys Arg
                    275                 280                 285
Glu Phe Arg Ala Asp Ala Arg Ala Val Ile Ser Leu Met Glu Met
                    290                 295                 300
Gly Phe Asp Glu Lys Glu Val Ile Asp Ala Leu Arg Val Asn Asn
                    305                 310                 315
Asn Gln Gln Asn Ala Ala Cys Glu Trp Leu Leu Gly Asp Arg Lys
                    320                 325                 330
Pro Ser Pro Glu Glu Leu Asp Lys Gly Ile Asp Pro Asp Ser Pro
                    335                 340                 345
Leu Phe Gln Ala Ile Leu Asp Asn Pro Val Val Gln Leu Gly Leu
                    350                 355                 360
Thr Asn Pro Lys Thr Leu Leu Ala Phe Glu Asp Met Leu Glu Asn
                    365                 370                 375
Pro Leu Asn Ser Thr Gln Trp Met Asn Asp Pro Glu Thr Gly Pro
                    380                 385                 390
Val Met Leu Gln Ile Ser Arg Ile Phe Gln Thr Leu Asn Arg Thr
                    395                 400                 405
```

<210> 9
<211> 177

<212> PRT
<213> Homo sapiens

<220>
<221> unsure
<222> 170, 171
<223> unknown, or other

<220> -
<223> 1871375

<400> 9
Met Val Met His Asn Ser Asp Pro Asn Leu His Leu Leu Ala Glu
1 5 10 15
Gly Ala Pro Ile Asp Trp Gly Glu Glu Tyr Ser Asn Ser Gly Gly
20 25 30
Gly Gly Ser Pro Ala Pro Ala Pro Arg Ser Gln Pro Pro Ser Arg
35 40 45
Lys Ser Asp Gly Ala Pro Ser Arg Trp Ser Leu Trp Ser Arg Met
50 55 60
Arg Arg Trp Gly Cys Pro Leu Arg Leu Ala Leu Ser His His His
65 70 75
Leu Arg Pro Arg Thr Val Ser Leu Arg Ser Glu Ala Cys Trp Pro
80 85 90
Lys Val Cys Gly Leu Arg Ala Pro His Gln Pro Ala Pro Cys Ser
95 100 105
Thr Gly Pro Pro Leu Gly Arg Val Pro Ser Leu Arg Pro Pro Pro
110 115 120
Arg Pro Pro Arg Arg Leu Pro His Pro Ser Ser Ile Ser Cys Leu
125 130 135
Glu Arg Leu Trp Thr Leu Gly Pro Pro Ser Pro Ala Thr Arg Arg
140 145 150
Leu Glu Ser Arg Cys Pro Ala Pro Ala Ala Thr Pro Pro Ser Thr
155 160 165
Pro Pro Pro Arg Xaa Xaa Phe Lys Gly Cys Lys Asn
170 175

<210> 10
<211> 197
<212> PRT
<213> Homo sapiens

<220> -
<223> 1880830

<400> 10
Met Ile Thr Cys Arg Val Cys Gln Ser Leu Ile Asn Val Glu Gly
1 5 10 15
Lys Met His Gln His Val Val Lys Cys Gly Val Cys Asn Glu Ala
20 25 30
Thr Pro Ile Lys Asn Ala Pro Pro Gly Lys Lys Tyr Val Arg Cys
35 40 45
Pro Cys Asn Cys Leu Leu Ile Cys Lys Val Thr Ser Gln Arg Ile
50 55 60
Ala Cys Pro Arg Pro Tyr Cys Lys Arg Ile Ile Asn Leu Gly Pro

```
                65                  70                  75
Val His Pro Gly Pro Leu Ser Pro Glu Pro Gln Pro Met Gly Val
                80                  85                  90
Arg Val Ile Cys Gly His Cys Lys Asn Thr Phe Leu Trp Thr Glu
                95                 100                 105
Phe Thr Asp Arg Thr Leu Ala Arg Cys Pro His Cys Arg Lys Val
               110                 115                 120
Ser Ser Ile Gly Arg Arg Tyr Pro Arg Lys Arg Cys Ile Cys Cys
               125                 130                 135
Phe Leu Leu Gly Leu Leu Leu Ala Val Thr Ala Thr Gly Leu Ala
               140                 145                 150
Phe Gly Thr Trp Lys His Ala Arg Arg Tyr Gly Gly Ile Tyr Ala
               155                 160                 165
Ala Trp Ala Phe Val Ile Leu Leu Ala Val Leu Cys Leu Gly Arg
               170                 175                 180
Ala Leu Tyr Trp Ala Cys Met Lys Val Ser His Pro Val Gln Asn
               185                 190                 195
Phe Ser
```

<210> 11
<211> 346
<212> PRT
<213> Homo sapiens

<220> -
<223> 2328134

<400> 11

```
Met Thr Pro Arg Thr Trp Trp Pro Arg Pro Ala Gly Trp Gly Thr
  1               5                  10                  15
Cys Arg Ala Ala Gly Trp Pro Arg Ser Val Pro Trp Ala Arg Thr
                20                  25                  30
Ala Ala Ser Leu Val Phe Val Pro Thr Arg Arg Arg Ser Gly Pro
                35                  40                  45
Ser Gly Thr Ala Ser Val Ala Ala Met Ala Tyr His Ser Gly Tyr
                50                  55                  60
Gly Ala His Gly Ser Lys His Arg Ala Arg Ala Ala Pro Asp Pro
                65                  70                  75
Pro Pro Leu Phe Asp Asp Thr Ser Gly Gly Tyr Ser Ser Gln Pro
                80                  85                  90
Gly Gly Tyr Pro Ala Thr Gly Ala Asp Val Ala Phe Ser Val Asn
                95                 100                 105
His Leu Leu Gly Asp Pro Met Ala Asn Val Ala Met Ala Tyr Gly
               110                 115                 120
Ser Ser Ile Ala Ser His Gly Lys Asp Met Val His Lys Glu Leu
               125                 130                 135
His Arg Phe Val Ser Val Ser Lys Leu Lys Tyr Phe Phe Ala Val
               140                 145                 150
Asp Thr Ala Tyr Val Ala Lys Lys Leu Gly Leu Leu Val Phe Pro
               155                 160                 165
Tyr Thr His Gln Asn Trp Glu Val Gln Tyr Ser Arg Asp Ala Pro
               170                 175                 180
Leu Pro Pro Arg Gln Asp Leu Asn Ala Pro Asp Leu Tyr Ile Pro
               185                 190                 195
Thr Met Ala Phe Ile Thr Tyr Val Leu Leu Ala Gly Met Ala Leu
```

```
                200                 205                 210
Gly Ile Gln Lys Arg Phe Ser Pro Glu Val Leu Gly Leu Cys Ala
                215                 220                 225
Ser Thr Ala Leu Val Trp Val Val Met Glu Val Leu Ala Leu Leu
                230                 235                 240
Leu Gly Leu Tyr Leu Ala Thr Val Arg Ser Asp Leu Ser Thr Phe
                245                 250                 255
His Leu Leu Ala Tyr Ser Gly Tyr Lys Tyr Val Gly Met Ile Leu
                260                 265                 270
Ser Val Leu Thr Gly Leu Leu Phe Gly Ser Asp Gly Tyr Tyr Val
                275                 280                 285
Ala Leu Ala Trp Thr Ser Ser Ala Leu Met Tyr Phe Ile Val Arg
                290                 295                 300
Ser Leu Arg Thr Ala Ala Leu Gly Pro Asp Ser Met Gly Gly Pro
                305                 310                 315
Val Pro Arg Gln Arg Leu Gln Leu Tyr Leu Thr Leu Gly Ala Ala
                320                 325                 330
Ala Phe Gln Pro Leu Ile Ile Tyr Trp Leu Thr Phe His Leu Val
                335                 340                 345
Arg
```

<210> 12
<211> 256
<212> PRT
<213> Homo sapiens

<220> -
<223> 2652271

<400> 12

```
Met Arg Pro Ala Ala Leu Arg Gly Ala Leu Leu Gly Cys Leu Cys
  1                 5                  10                  15
Leu Ala Leu Leu Cys Leu Gly Gly Ala Asp Lys Arg Leu Arg Asp
                 20                  25                  30
Asn His Glu Trp Lys Lys Leu Ile Met Val Gln His Trp Pro Glu
                 35                  40                  45
Thr Val Cys Glu Lys Ile Gln Asn Asp Cys Arg Asp Pro Pro Asp
                 50                  55                  60
Tyr Trp Thr Ile His Gly Leu Trp Pro Asp Lys Ser Glu Gly Cys
                 65                  70                  75
Asn Arg Ser Trp Pro Phe Asn Leu Glu Glu Ile Lys Asp Leu Leu
                 80                  85                  90
Pro Glu Met Arg Ala Tyr Trp Pro Asp Val Ile His Ser Phe Pro
                 95                 100                 105
Asn Arg Ser Arg Phe Trp Lys His Glu Trp Glu Lys His Gly Thr
                110                 115                 120
Cys Ala Ala Gln Val Asp Ala Leu Asn Ser Gln Lys Lys Tyr Phe
                125                 130                 135
Gly Arg Ser Leu Glu Leu Tyr Arg Glu Leu Asp Leu Asn Ser Val
                140                 145                 150
Leu Leu Lys Leu Gly Ile Lys Pro Ser Ile Asn Tyr Tyr Gln Val
                155                 160                 165
Ala Asp Phe Lys Asp Ala Leu Ala Arg Val Tyr Gly Val Ile Pro
                170                 175                 180
```

```
Lys Ile Gln Cys Leu Pro Pro Ser Gln Asp Glu Glu Val Gln Thr
                185                 190                 195
Ile Gly Gln Ile Glu Leu Cys Leu Thr Lys Gln Asp Gln Gln Leu
            200                 205                 210
Gln Asn Cys Thr Glu Pro Gly Glu Gln Pro Ser Pro Lys Gln Glu
        215                 220                 225
Val Trp Leu Ala Asn Gly Ala Ala Glu Ser Arg Gly Leu Arg Val
    230                 235                 240
Cys Glu Asp Gly Pro Val Phe Tyr Pro Pro Pro Lys Lys Thr Lys
245                 250                 255
His
```

<210> 13
<211> 235
<212> PRT
<213> Homo sapiens

<220> -
<223> 2965248

<400> 13

```
Met Ala Ser Thr Ile Ser Ala Tyr Lys Glu Lys Met Lys Glu Leu
  1               5                  10                  15
Ser Val Leu Ser Leu Ile Cys Ser Cys Phe Tyr Thr Gln Pro His
            20                  25                  30
Pro Asn Thr Val Tyr Gln Tyr Gly Asp Met Glu Val Lys Gln Leu
        35                  40                  45
Asp Lys Arg Ala Ser Gly Gln Ser Phe Glu Val Ile Leu Lys Ser
    50                  55                  60
Pro Ser Asp Leu Ser Pro Glu Ser Pro Met Leu Ser Ser Pro Pro
65                  70                  75
Lys Lys Lys Asp Thr Ser Leu Glu Glu Leu Gln Lys Arg Leu Glu
                80                  85                  90
Ala Ala Glu Glu Arg Arg Lys Thr Gln Glu Ala Gln Val Leu Lys
            95                  100                 105
Gln Leu Ala Asp Gly Ala Ser Thr Ser Ala Arg Cys Cys Thr Arg
        110                 115                 120
Arg Trp Arg Arg Ile Thr Thr Ser Ala Ala Arg Arg Arg Arg Ser
    125                 130                 135
Ser Thr Thr Arg Trp Ser Ser Ala Arg Arg Ser Ala Arg His Thr
140                 145                 150
Trp Pro His Cys Ala Ser Gly Cys Ala Arg Arg Ser Cys Thr Arg
            155                 160                 165
Pro Arg Cys Ala Gly Thr Arg Ser Ser Glu Lys Arg Cys Arg Ala
        170                 175                 180
Lys Gly Pro Gly Arg Ala Ala Pro Ile Leu Arg Arg Asn Thr Phe
    185                 190                 195
Gly Phe Trp Phe Cys Phe Val His Leu Cys Leu Asp Ala Thr Phe
200                 205                 210
Val Pro Pro Pro Pro Pro Gln Pro Pro Ala Ser Cys Phe Ser Ser
            215                 220                 225
Ala Leu Ser Arg Pro Ala Leu Ser Ser Trp
        230                 235
```

<210> 14
<211> 371
<212> PRT
<213> Homo sapiens

<220> -
<223> 3057669

<400> 14

```
Met Asp His Glu Asp Ile Ser Glu Ser Val Asp Ala Ala Tyr Asn
 1               5                  10                  15
Leu Gln Asp Ser Cys Leu Thr Asp Cys Asp Val Glu Asp Gly Thr
            20                  25                  30
Met Asp Gly Asn Asp Glu Gly His Ser Phe Glu Leu Cys Pro Ser
        35                  40                  45
Glu Ala Ser Pro Tyr Val Arg Ser Arg Glu Arg Thr Ser Ser Ser
    50                  55                  60
Ile Val Phe Glu Asp Ser Gly Cys Asp Asn Ala Ser Ser Lys Glu
65                  70                  75
Glu Pro Lys Thr Asn Arg Leu His Ile Gly Asn His Cys Ala Asn
                80                  85                  90
Lys Leu Thr Ala Phe Lys Pro Thr Ser Ser Lys Ser Ser Ser Glu
            95                  100                 105
Ala Thr Leu Ser Ile Ser Pro Pro Arg Pro Thr Thr Leu Ser Leu
        110                 115                 120

Asp Leu Thr Lys Asn Thr Thr Glu Lys Leu Gln Pro Ser Ser Pro
    125                 130                 135
Lys Val Tyr Leu Tyr Ile Gln Met Gln Leu Cys Arg Lys Glu Asn
140                 145                 150
Leu Lys Asp Trp Met Asn Gly Arg Cys Thr Ile Glu Glu Arg Glu
155                 160                 165
Arg Ser Val Cys Leu His Ile Phe Leu Gln Ile Ala Glu Ala Val
            170                 175                 180
Glu Phe Leu His Ser Lys Gly Leu Met His Arg Asp Leu Lys Pro
        185                 190                 195
Ser Asn Ile Phe Phe Thr Met Asp Asp Val Val Lys Val Gly Asp
    200                 205                 210
Phe Gly Leu Val Thr Ala Met Asp Gln Asp Glu Glu Glu Gln Thr
215                 220                 225
Val Leu Thr Pro Met Pro Ala Tyr Ala Arg His Thr Gly Gln Val
            230                 235                 240
Gly Thr Lys Leu Tyr Met Ser Pro Glu Gln Ile His Gly Asn Ser
        245                 250                 255
Tyr Ser His Lys Val Asp Ile Phe Ser Leu Gly Leu Ile Leu Phe
    260                 265                 270
Glu Leu Leu Tyr Pro Phe Ser Thr Gln Met Glu Arg Val Arg Thr
275                 280                 285
Leu Thr Asp Val Arg Asn Leu Lys Phe Pro Pro Leu Phe Thr Gln
            290                 295                 300
Lys Tyr Pro Cys Glu Tyr Val Met Val Gln Asp Met Leu Ser Pro
        305                 310                 315
Ser Pro Met Glu Arg Pro Glu Ala Ile Asn Ile Ile Glu Asn Ala
    320                 325                 330
Val Phe Glu Asp Leu Asp Phe Pro Gly Lys Thr Val Leu Arg Gln
335                 340                 345
```

```
Arg Ser Arg Ser Leu Ser Ser Ser Gly Thr Lys His Ser Arg Gln
                350                 355                 360
Ser Asn Asn Ser His Ser Pro Leu Pro Ser Asn
                365                 370
```

<210> 15
<211> 523
<212> PRT
<213> Homo sapiens

<220> -
<223> 3125156

<400> 15

```
Met Gly Pro Gln Ala Ala Pro Leu Thr Ile Arg Gly Pro Ser Ser
  1               5                  10                  15
Ala Gly Gln Ser Thr Pro Ser Pro His Leu Val Pro Ser Pro Ala
                20                  25                  30
Pro Ser Pro Gly Pro Gly Pro Val Pro Pro Arg Pro Pro Ala Ala
                35                  40                  45
Glu Pro Pro Pro Cys Leu Arg Arg Gly Ala Ala Ala Ala Asp Leu
                50                  55                  60
Leu Ser Ser Ser Pro Glu Ser Gln His Gly Gly Thr Gln Ser Pro
                65                  70                  75
Gly Gly Gly Gln Pro Leu Leu Gln Pro Thr Lys Val Asp Ala Ala
                80                  85                  90

Glu Gly Arg Arg Pro Gln Ala Leu Arg Leu Ile Glu Arg Asp Pro
                95                 100                 105
Tyr Glu His Pro Glu Arg Leu Arg Gln Leu Gln Gln Glu Leu Glu
               110                 115                 120
Ala Phe Arg Gly Gln Leu Gly Asp Val Gly Ala Leu Asp Thr Val
               125                 130                 135
Trp Arg Glu Leu Gln Asp Ala Gln Glu His Asp Ala Arg Gly Arg
               140                 145                 150
Ser Ile Ala Ile Ala Arg Cys Tyr Ser Leu Lys Asn Arg His Gln
               155                 160                 165
Asp Val Met Pro Tyr Asp Ser Asn Arg Val Val Leu Arg Ser Gly
               170                 175                 180
Lys Asp Asp Tyr Ile Asn Ala Ser Cys Val Glu Gly Leu Ser Pro
               185                 190                 195
Tyr Cys Pro Pro Leu Val Ala Thr Gln Ala Pro Leu Pro Gly Thr
               200                 205                 210
Ala Ala Asp Phe Trp Leu Met Val His Glu Gln Lys Val Ser Val
               215                 220                 225
Ile Val Met Leu Val Ser Glu Ala Glu Met Glu Lys Gln Lys Val
               230                 235                 240
Ala Arg Tyr Phe Pro Thr Glu Arg Gly Gln Pro Met Val His Gly
               245                 250                 255
Ala Leu Ser Leu Ala Leu Ser Ser Val Arg Ser Thr Glu Thr His
               260                 265                 270
Val Glu Arg Val Leu Ser Leu Gln Phe Arg Asp Gln Ser Leu Lys
               275                 280                 285
Arg Ser Leu Val His Leu His Phe Pro Thr Trp Pro Glu Leu Gly
               290                 295                 300
```

```
Leu Pro Asp Ser Pro Ser Asn Leu Leu Arg Phe Ile Gln Glu Val
                305                 310                 315
His Ala His Tyr Leu His Gln Arg Pro Leu His Thr Pro Ile Ile
                320                 325                 330
Val His Cys Ser Ser Gly Val Gly Arg Thr Gly Ala Phe Ala Leu
                335                 340                 345
Leu Tyr Ala Ala Val Gln Glu Val Glu Ala Gly Asn Gly Ile Pro
                350                 355                 360
Glu Leu Pro Gln Leu Val Arg Arg Met Arg Gln Gln Arg Lys His
                365                 370                 375
Met Leu Gln Glu Lys Leu His Leu Arg Phe Cys Tyr Glu Ala Val
                380                 385                 390
Val Arg His Val Glu Gln Val Leu Gln Arg His Gly Val Pro Pro
                395                 400                 405
Pro Cys Lys Pro Leu Ala Ser Ala Ser Ile Ser Gln Lys Asn His
                410                 415                 420
Leu Pro Gln Asp Ser Gln Asp Leu Val Leu Gly Gly Asp Val Pro
                425                 430                 435
Ile Ser Ser Ile Gln Ala Thr Ile Ala Lys Leu Ser Ile Arg Pro
                440                 445                 450
Pro Gly Gly Leu Glu Ser Pro Val Ala Ser Leu Pro Gly Pro Ala
                455                 460                 465
Glu Pro Pro Gly Leu Pro Pro Ala Ser Leu Pro Glu Ser Thr Pro
                470                 475                 480
Ile Pro Ser Ser Ser Gln Thr Pro Phe Pro Pro His Tyr Leu Arg
                485                 490                 495
Leu Pro Ser Leu Arg Arg Ser Arg Gln Cys Leu Lys Pro Pro Ala
                500                 505                 510
Arg Gly Pro Pro Pro Pro Pro Trp Asn Cys Trp Pro Pro
                515                 520
```

<210> 16
<211> 846
<212> DNA
<213> Homo sapiens

<220> -
<223> 866885

<400> 16

```
ggcgggcgga gtctgcagga tggcaccgga cccctggttc tccacatacg attctacttg  60
tcaaattgcc caagaaattg ctgagaaaat tcaacaacga aatcaatatg aaagaaaagg 120
tgaaaaggca ccaaagctta ccgtgacaat cagagctttg ttgcagaacc tgaaggaaaa 180
gatcgccctt ttgaaggact tattgctaag agctgtgtca acacatcaga taacacagct 240
tgaaggggac cgaagacaga acctcttgga tgatcttgta actcgagaga gactacttct 300
ggcatccttt aagaatgagg gtgccgaacc agatctaatc aggtccagcc tgatgagtga 360
agaggctaag cgaggagcac ccaacccttg gctctttgag gagccagagg agaccagagg 420
cttgggtttt gatgaaatcc ggcaacagca gcagaaaatt atccaagaac aggatgcagg 480
ccttgatgcc ctttcctcta tcataagtcg ccaaaaacaa atggggcagg aaattgggaa 540
tgaattggat gaacaaaatg agataattga cgaccttgcc aacctagtgg agaacacaga 600
tgaaaaactt cgcaatgaaa ccaggcgggt aaacatggtg gacagaaagt cagcctcttg 660
tgggatgatc atggtgattt tactgctgct tgtggctatc gtggttgttg cagtctggcc 720
gaccaactga tggcagtaaa gagaccacca gcagtgacac ctggcaatga cagatgcaag 780
cccaacaccc ttttggtacg caaaacctgc tctcaataaa ttcccccaaa gctctgaaaa 840
aaaaaa                                                          846
```

<210> 17
<211> 1897
<212> DNA
<213> Homo sapiens

<220> -
<223> 1273453

<400> 17
tgcacatcta gcacaaattg aagatgatag agctgcgatg gttatttctt ggcatctggc 60
aagtgacatg gactgtgtag tcaccctaac cactgacgct gcacgtcgta tctatgatga 120
aacccaaggt cgtcagcagg tgttgcccct tgattctatt tacaagaaga ctcttccaga 180
ttggaaaaga tctctacctc atttccgaaa tggaaaattg tattttaaac ccattggaga 240
tccagtcttt gctcgagact tgttaacatt tccagataat gtagaacatt gtgaaacagt 300
atttggtatg ctgttaggag acaccattat tttggataat ctggatgcgg ccaatcatta 360
tagaaaagag gttgttaaaa ttacacactg tcctacactg ctgaccagag atggagatcg 420
aattcgaagt aatggaaagt ttgggggcct tcagaataaa gctcctccaa tggataaact 480
tcggggaatg gtatttggag ctccagttcc aaaacagtgt ctgatcttag gggaacaaat 540
agatcttctt cagcagtatc gttctgctgt gtgcaaacta gacagtgtga ataaggatct 600
taacagtcaa ttagagtacc ttcgcactcc ggatatgagg aagaaaaagc aagaacttga 660
tgaacatgag aaaaatctca aactaataga ggaaaaacta ggtatgactc ccatacgtaa 720
gtgtaatgac tcattgcgtc attcaccaaa ggttgagacg acagattgtc cagttcctcc 780
taaaagaatg agacgagaag ctacaagaca aaataggatt ataaccaaaa cagatgtatg 840
agaggtgaca gagagaagag gccattggtc tcagtaagaa tgccctgctt tctgcatctc 900
tgtttcagaa gaccaagagg gtgacttacc agactgagta tttctgggga caatacaagt 960
acctgggcat gaatttccat ttcgattcag atgggactgg aaacaaccat tcaattttat 1020
gaatcttact ggacattatg gatttactgg aattattcca gacattatgc cctttggttg 1080
tcactacctt gcaaatgtgt aagaggaaaa tgtgctaatg tggcagtgac tgtaaaactg 1140
gcacatggca tttattaatc ctgaagaaaa gtacatgtac tatttttcag tataaatata 1200
atgaacatgt cagaactatt tcttgaaaac ctttttatta cttttgcgtg aatttattta 1260
acaaagatgt tttgtctttt gtgtaaggga ggttctagag gctagatgtt taattgtaaa 1320
tatgtgagga aactcaatgc agaattcagg ataaaaattt taaaagcaca ggtatttggg 1380
aattgaaatg ttaagatacc cagaacaaca ttaaatcaat gagtgaactt gtgacagtgg 1440
tagcatttca aatttcaaaa gacttatcct gtgtgtgtgt gtgtgtgtgt atatatatat 1500
atatatatat aaatatatat atataaaata ttcagcagca ccaagtttta taactattgt 1560
ttgtttgact ttattaatac tagaatatgt agtctcagcc ttaattttac atttacatta 1620
ttttgtaatt ttttattact atttttaagg ggttaaagag aacatacatt ctcacattag 1680
tgtactttct ggtagaaagt tgctgcaaaa acatttgaaa tgtatattaa cctaatgtat 1740
gtcatatata tgtctttgtg taagttcaag actattgatc tgtgaagtta ttttgtaagg 1800
acatacattt ggtaagtaag tttgtgtccc aggaaatgta tgtgtttttа aaccctttct 1860
aaatatgcag gccattaata aataagattg tgtctca 1897

<210> 18
<211> 2272
<212> DNA
<213> Homo sapiens

<220> -
<223> 1534876

<400> 18
ccatgctcca ggcatacaga tgtggtttct cggctgcacc gggccaggct gcgggtgtgc 60
aggcgtctgc aaagttgtgc catgtatcag cacaggcttt gagacgtctg gaccctgtcc 120
ttcctcccgt gaggggttct tgttctttct gactcaggtg acttttcagc ccttccaatt 180
cccctctttt tctgccctcc cctccaactc agccaaccca ggtgtgggca gtcagggagg 240

gagggagtgt cccaccacgt tctcagggca gcccttgact cctaagcccc ttcctccttc 300
cattctgcat ccccteccca tccaacctaa atgcccacag ctggggctga gctgtattcc 360
tgtggaggga cctctgccgt gcctctctga ggtcaggctg tgctgtgtga tgggcaggct 420
ttgccccagc ccaccctgg caaggtgcac ttgttttctg gtttgtacaa ggtgtcctgg 480
gggcccgtcg cttccctgcc agtgaggagt gacttctccc tctcttccag tcctgtaggg 540
gagacaaaac cagattgggg ggcccaaggg gagcatggaa aaggccggct cccctgtctt 600
tccttggctg tcagagtcag ggtaacacac accaagagtg gagtgcggcc agcaagtttg 660
agacctgccc gccctcctcg cagctctgct ctgtgtcctc aggaagtcac agagtctact 720
gaggcaagga gagggtgatt ctttccccaa atcccttctt ccctggttcc caaaccaaag 780
acagcctgca gcccttttctg catggggtgc tctgttgaca ggcttcccag atccctgagt 840
ctctctttcc ttcctcctcg atctttagtt gtccacggtc aattcagtgc ttccattggg 900
ggacagtccc ctccgggatg acctgattca cctccagccc agggaatgga atctagagga 960
atacgtgggg tgggtctgga caaggagcgg caggaatcac cacccatctc cagctgtgga 1020
gccctgtgga ggggaagggg aagcttgggg ttcagaggga actcttccag gagaggggtg 1080
cccagcggag gtaaagatga tagagggttg tgggggggtct ctagttgaat gttttggccc 1140
atgactttgg aacatggctg gcagcttcca gcagaagtca cgctccccat cccccagggg 1200
acataggacc ttttttcctgc ttcctggtca ctttcaaaga actatttgcg caatctgtgg 1260
gtctgtggat tcacggggct ttctgtgtgg gtgctgcagt tgcttttgtc tgcagcagca 1320
ggacacatct ttcctcttac tcagcccttt atggcccatg gggaactccg tggctcaggg 1380
agagctgaac tccaggggtg tgacctggga caggtgggcc tgaggtgccc agctcagggc 1440
agccaggtgg ctcatgggct gtagtgagcc agctccctgg gggaaaaggc tgtgggccgt 1500
taggaccatc ctccaggaca ggtgacctct atgaggtcac ctacggctgt ggccgtgcag 1560
gcctccttcc agcccagagt ggcccagtag agcaaggcag acagtgacct ccacccccgc 1620
agccctctta aaaggccagt actcttgggg gtgggggggag ggtttagaaa gcatttgccc 1680
atctgccttt ctttcccccа gcccccaccc gctttgaatg tagagacccg tgggcacttt 1740
tccttttgtg gtgggggggtg cggaggaggt acccccaccc ctggcacagc cgcctggaat 1800
gcaggactgt cactgctgtt cgggtgatga cctcgttgcc aagctcctcc tgtccccttg 1860
ttctgggggc aggcgctgtg cttctgtgag gtggtttagc ttttgctttc gaagtggcca 1920
gctgcggcca ccaggtctca gcacaagagc gcttcctttg cacagaatga gcttcgagct 1980
ttgttcagac taaatgaatg tatctgggag gggtcggggg cacgagttga ttccaagcac 2040
atgcctttgc tgagtgtgtg tgtgctggga gagtcagagt ggatgtagag cgcggtttta 2100
tttttgtact gacattggta agagactgta tagcatctat ttatttagat gatttatctg 2160
gtaaatgagg caaaaaaatt attaaaaata cattaaagat gatttaaaaa aaagaccaaa 2220
aaaccaagaa acccaaagcc caagaatgcg cgtagcatcc aaaaaaaaaa gg 2272

<210> 19
<211> 992
<212> DNA
<213> Homo sapiens

<220> -
<223> 1634813

<400> 19
gacagcttgg cctacagccc ggcgggcatc agctcccttg acccagtgga tatcggtggc 60
cccgttattc gtccaggtgc ccagggagga ggacccgcct gcagcatgaa cctgtggctc 120
ctggcctgcc tggtggccgg cttcctggga gcctgggccc ccgctgtcca cgcccaaggt 180
gtctttgagg actgctgcct ggcctaccac taccccattg ggtgggctgt gctccggcgc 240
gcctggactt accggatcca ggaggtgagc gggagctgca atctgcctgc tgcgatattc 300
tacctccccа agagacacag gaaggtgtgt gggaacccca aaagcaggga ggtgcagaga 360
gccatgaagc tcctggatgc tcgaaataag gtttttgcaa agctccgcca caacacgcag 420
accttccaag caggccctca tgctgtaaag aagttgagtt ctggaaactc caagttatca 480
tcatccaagt ttagcaatcc catcagcagc agcaagagga atgtctccct cctgatatca 540
gctaattcag gactgtgagc cggctcattt ctgggctcca tcggcacagg aggggccgga 600
tctttctccg ataaaaccgt cgccctacag acccagctgt ccccacgcct ctgtcttttg 660

ggtcaagtct taatccctgc acctgagttg gtcctccctc tgcaccccca ccacctcctg 720
cccgtctggc aactggaaag agggagttgg cctgatttta agccttttgc cgctccgggg 780
accagcagca atcctgggca gccagtggct cttgtagaga agacttagga tacctctctc 840
actttctgtt tcttgccgtc caccccgggc catgccagtg tgtccctctg ggtccctcca 900
aaactctggt cagttcaagg atgcccctcc caggctatgc ttttctataa cttttaaata 960
aaccttgggg gttgatggag tcaaaaaaaa aa 992

<210> 20
<211> 810
<212> DNA
<213> Homo sapiens

<220> -
<223> 1711840

<400> 20
cgagtgagcg cgcggcggcc cctggtccgc ccggccgcgg ccgatctagg ggctgggggc 60
tggaggcggg ggtgggggtc tgagctgcgt cctgggctcg aggcgtcccc cggggagtcg 120
cctcttagcg gtgcgtccgg gctagcggcg aggggccgcc ccaagtcttc ccaccgccgc 180
caccttagca gcccgacttg gggcctggaa agtggagcac gcggaggtgg gagggccctg 240
cacgcggccc ccggtgggga aggggacggg ccagggattc agactcgggc tctcccctca 300
ggatgcagca ccgaggcttc ctcctcctca ccctcctcgc cctgctggcg ctcacctccg 360
cggtcgccaa aaagcaagat aaggtgaaga agggcggccc ggggagcgag tgcgctgagt 420
gggcctgggg gccctgcacc cccagcagca aaggatttgc ggcagtgggt tttccgcgag 480
ggccaccttg gggggggccca agaacccaac cggcagtcct ggttgaaagg gttgcccctg 540
gaaagttgga aagaaaggag ttttgggcac ccggactttg gaaagttggc caaatttttt 600
ggaagaaaac ttggcgggtc tgccggtccg ttaaatgggg gaggggacaa aagaattgaa 660
agccgaaaaa atgctttctc cgccgccaag agaggtcgaa cccgcgtctg gcaagaagag 720
aaaagggcgc gcccacactg ttaacaacaa tatggcgcct gaacagttgg tggcaccaca 780
gggggaggga gacacatact tgcgcgcggt 810

<210> 21
<211> 1064
<212> DNA
<213> Homo sapiens

<220> -
<223> 1747327

<400> 21
ttcctggggc tccggggcgc ggagaagctg catcccagag gagcgcgtcc aggagcggac 60
ccgggagtgt ttcaagagcc agtgacaagg accaggggcc caagtcccac cagccatgca 120
gacctgcccc ctggcattcc ctggccacgt ttcccaggcc cttgggaccc tcctgttttt 180
ggctgcctcc ttgagtgctc agaatgaagg ctgggacagc cccatctgca cagaggggt 240
agtctctgtg tcttggggcg agaacaccgt catgtcctgc aacatctcca acgccttctc 300
ccatgtcaac atcaagctgc gtgcccacgg gcaggagagc gccatcttca atgaggtggc 360
tccaggctac ttctcccggg acggctggca gctccaggtt cagggaggcg tggcacagct 420
ggtgatcaaa ggcgcccggg actcccatgc tgggctgtac atgtggcacc tcgtgggaca 480
ccagagaaat aacagacaag tcacgctgga ggtttcaggt gcagaacccc agtccgcccc 540
cgacactggg ttctggcctg tgccagcggt ggtcactgct gtcttcatcc tcttggtcgc 600
tctggtcatg ttcgcctggt acaggtgccg ctgttcccag caacgccggg agaagaagtt 660
cttcctccta gaaccccaga tgaaggtcgc agccctcaga gcgggagccc agcagggcct 720
gagcagagcc tccgctgaac tgtggacccc agactccgag cccaccccaa ggccgctggc 780
actggtgttc aaaccctcac cacttggagc cctggagctg ctgtcccccc aacccttgtt 840

tccatatgcc gcagacccat agccgcctgc aaggaagaga ggacacagga gtagccaccc 900
tgagtgccga cctttggtgg cgggggcctg ggtctctcgt ccccacccgg aagggcacaa 960
gacaccgggc tttgcttggc aaggcttggg gcctcttgtg gtcaacccag ttcccttggg 1020
tgccgttgca gaacccctta gcccctttcca acgtcgacca ggtt 1064

<210> 22
<211> 1336
<212> DNA
<213> Homo sapiens

<220> -
<223> 1864292

<400> 22
agctcgtacc cctcgagtga aattctgaaa tgaagatgga ggaggcagtg ggaaaagttg 60
aagaactcat tgagtccgaa gccccaccaa aagcatctga acaagagaca gccaaggagg 120
aagatggatc tgtagaactg gaatctcaag ttcagaaaga tggtgtagcg gattctacag 180
ttatttcttc aatgccctgc ttgttgatgg aactgagaag ggactcttct gagtctcagt 240
tagcatccac agagagtgac aagcctacaa ctggccgagt ttatgagagt gacccctcta 300
atcactgcat gctttccccct tcctctagtg gtcacctggc tgattcagat acgttgtctt 360
ccgcagaaga gaatgaaccc tctcaggcag aaacggcggt agaaggagac ccttcaggag 420
tgtctggtgc cacagttggg cgcaagtcta ggcggtcccg atctgaaagt gaaacttcca 480
ctatggctgc caagaaaaac cggcaatcca gtgataaaca gaatggccga gtcgccaagg 540
ttaaaggtca tcggagccaa aagcacaagg agaggatcag gctactgagg cagaaacggg 600
aggctgctgc aaggaagaaa tataacctgc tgcaggacag tagtaccagt gatagtgacc 660
tgacttgtga ctcaagcacg agctcatcag atgatgatga agaggtttca gggagcagca 720
agacaatcac tgcagagata ccagatggac ctccagttgt agctcattat gatatgtctg 780
acaccaactc tgacccagaa gtggtaaatg tggacaattt attggcggct gcagtagttc 840
aagagcacag taattctgta ggcggccagg acacaggagc tacctggagg accagcgggc 900
ttctagagga gctgaatgca gaggcaggtc atttggatcc aggattccta gcaagtgaca 960
aaacatctgc tggcaatgcg ccactcaatg aagaaattaa cattgcgtct tcagatagtg 1020
aagtagagat tgtgggagtt caggaacatg caaggtgtgt tcatcctcga ggtggtgtga 1080
ttcagagtgt ttcttcatgg aagcatggct cgggcacgca gtatgttagc accaggcaaa 1140
cacagtcatg gactgctgtg actccccagc agacttgggc ttcaccagca gaagttgttg 1200
accttacctt ggatgaggat agcaggcgta aatacctact gtaatacaat gtcactgtgt 1260
ttcctctgca ctgttccctt ccacttcctc atcctctttg tgacatggaa gttcattgtc 1320
ataggggtac ggagct 1336

<210> 23
<211> 1742
<212> DNA
<213> Homo sapiens

<220> -
<223> 1866437

<400> 23
gccccgcccc ctccccgccc gccttcccgg tgaccttcag gggcccgggt ggcgggcgca 60
ggcccctgcg gcggcggcgg gatgttcgtg caggaggaga agatcttcgc gggcaaggtg 120
ctgcggctgc acatctgcgc gtccgacggc gccgagtggc tggaggaggc caccgaggac 180
acctcggtgg agaagctcaa ggagcgctgc ctcaagcact gtgctcatgg gagcttagaa 240
gatcccaaaa gtataaccca tcataaatta atccacgctg cctcagagag ggtgctgagt 300
gatgccagga ccatcctgga agagaacatc caggaccaag atgtcctatt attgaaaaaa 360
aagcgtgctc catcaccact tcccaagatg gctgatgtct cagcagaaga aaagaaaaaa 420

caagaccaga aagctccaga taaagaggcc atactgcggg ccaccgccaa cctgccctcc 480
tacaacatgg accgggccgc ggtccagacc aacatgagag acttccagac agaactccgg 540
aagatactgg tgtctctcat cgaggtggcg cagaagctgt tagcgctgaa cccagatgcg 600
gtggaattgt ttaagaaggc gaatgcaatg ctggacgagg acgaggatga gcgtgtggac 660
gaggctgccc tgcggcagct cacggagatg ggctttccgg agaacagagc caccaaggcc 720
cttcagctga accacatgtc ggtgcctcag gccatggagt ggctaattga acacgcagaa 780
gacccgacca tagacacgcc tcttcctggc caagctcccc cagaggccga gggggccaca 840
gcagctgcct ccgaggctgc cgcgggagcc agcgccaccg atgaggaggc cagagatgag 900
ctgacggaaa tcttcaagaa gatccggagg aaaagggagt ttcgggctga tgctcgggcc 960
gtcatttccc tgatggagat ggggttcgac gagaaagagg tgatagatgc cctcagagtg 1020
aacaacaacc agcagaatgc cgcgtgcgag tggctgctgg gggaccggaa gccctctccg 1080
gaggagctgg acaagggcat cgaccccgac agtcctctct ttcaggccat cctggataac 1140
ccggtggtgc agctgggcct gaccaacccg aaaacattgc tagcatttga agacatgctg 1200
gagaacccac tgaacagcac ccagtggatg aatgatccag aaacggggcc tgtcatgctg 1260
cagatctcta gaatcttcca gacactaaat cgcacgtagg tggcgttgtt ccactcggct 1320
atcaggccac agcagccccc tggtgcggcc cgagaccggg cagagtggac ctcacctgga 1380
aactcacctt cagcgcctca gccctggact gttagaggtg ctgcagctgc tcctgctctc 1440
tgatcttatt gcttataaac tttggtgacg gtagtgtgta aggccgtatt tttagcatct 1500
gacaggtgtt tacaaaaaag tggttgtcgc actgggaagt ggagtgatgg cctcgtctcc 1560
agtgctcctc tgggctcttg agttgctgct tgaattgccg tgtagacatt tgcttggaga 1620
gtccacttgt tatttgacgg aggtaggttt caacccagag ttaatgtcaa gcatgctaat 1680
ttaactagtc actcacagat gacttttctt taataaagtc ccttttccta ttaaaaaaaa 1740
aa 1742

<210> 24
<211> 1074
<212> DNA
<213> Homo sapiens

<220>
<221> unsure
<222> 546, 548, 573, 577, 588, 610, 615, 618, 647, 658, 676, 690,
<221> unsure
<222> 708, 718, 730, 738, 741, 770, 775, 780, 783, 803, 813, 818,
<221> unsure
<222> 819, 839, 844, 864, 870, 890, 901, 903, 918, 921, 943, 959,
<221> unsure
<222> 960, 967, 975, 990, 997, 999, 1000, 1007, 1040, 1053, 1055,
<221> unsure
<222> 1065, 1071
<223> a or g or c or t, unknown, or other

<220> -
<223> 1871375

<400> 24
gcggtgcaga ggaagcacaa cctctaccgg gacagcatgg tcatgcacaa cagcgacccc 60
aacctgcacc tgctggccga gggcgccccc atcgactggg gcgaggagta cagcaacagc 120
ggcgggggcg gcagcccagc cccagcaccc cggagtcagc caccctctcg gaaaagcgac 180
ggcgcgccaa gcaggtggtc tctgtggtcc aggatgagga ggtggggctg ccctttgagg 240
ctagccctga gtcaccacca cctgcgtccc cggacggtgt cactgagatc cgaggcctgc 300
tggcccaagg tctgcggcct gagagccccc caccagccgg ccccctgctc aacggggccc 360
ccgctgggga gagtccccag cctaaggccg ccccgaggc ctcctcgccg cctgcctcac 420
ccctccagca tctcctgcct ggaaaggctg tggaccttgg gccccccaag cccagcgacc 480
aggagactgg agagcaggtg tccagccccа gcagccaccc cgcctccac accaccaccg 540

aggacnantt tcaaggggtg caagaattga agnttcntaa gggccaantt gggggtcccc 600
ttgacttggn ttggnaanat tggggcaaaa agggccggtt ttccccnttt cccggganac 660
cccaagggaa aggggnttca aagcttcttn gggggggaaa gggggaancc cttgggtntt 720
ttgttggccn tttgtganca ncagcgagga gagtgcaaag gtgcagagtn agttntaggn 780
cantgggtcc ctgactgctg canatggtaa ggncgttnnc ttgtggaccc aaggcaggna 840
aagntgtggg gagggaagct ggtntgtgcn ttgtgggtgg aagcggggan ggctgtgttg 900
nanggcaggg agagggcnaa ntgagttatt tattggggtt cangtgaaaa gtttcttgnn 960
ccctgtnttg tgttnctgtg ggattgattn taagatngnn aggggtnggt ttttggggtt 1020
ttcctggttg gtggccaaan gggttggaaa atngntgggg ggggnttgga naat 1074

<210> 25
<211> 1454
<212> DNA
<213> Homo sapiens

<220> -
<223> 1880830

<400> 25
cccgggggag gcctgacccc ctccgcacca ccgtacggag ccgcatttcc cccgtttccc 60
gaggggcatc cagccgtgtt gcctggggag gacccacccc cctattcacc cttaactagc 120
ccggacagtg ggagtgcccc tatgatcacc tgccgagtct gccaatctct catcaacgtg 180
gaaggcaaga tgcatcagca tgtagtcaaa tgtggtgtct gcaatgaagc caccccaatc 240
aagaatgcac ccccagggaa aaaatatgtt cgatgccccc gtaactgtct ccttatctgc 300
aaagtgacat cccaacggat tgcatgcccc cggccctact gcaaaagaat catcaacctg 360
gggcctgtgc atcccggacc tctgagtcca gaaccccaac ccatgggtgt cagggttatc 420
tgtggacatt gcaagaatac ttttctgtgg acagagttca cagaccgcac tttggcacgt 480
tgtcctcact gcaggaaagt gtcatctatt gggcgcagat acccacgtaa gagatgtatc 540
tgctgcttct tgcttggctt gcttttggca gtcactgcca ctggccttgc ctttggcaca 600
tggaagcatg cacggcgata tggaggcatc tatgcagcct gggcatttgt catcctgttg 660
gctgtgctgt gtttgggccg ggctctttat tgggcctgta tgaaggtcag ccaccctgtc 720
cagaacttct cctgagcctg atgacccaca gactgtgcct ggcccctccc tggtggggac 780
agtgacacta cgaagggagc tggggtagtt aaaggctccc ggggcttcta gaaggaagcc 840
aagcagctgc cttcctttтc cctggggaga ggtaggaagg aaccaggccc tcacttaggt 900
ttggaggggc agataagagc actgctgacc atctgctttc ctccaagggt tgctgtgtct 960
agggtgaagt aggcaaaacg ttgcccttaa aactgggccc tgaagacggt tccagccttg 1020
tccttcctgt gtgctccctg agagccattc ctgtccctta cacattccag ggcagggtgg 1080
gggtgggtag ccctgggggt tcccctcccc cttgtgcacc attaggactt tgctgctgct 1140
attgcacttc accagaggtt ggctctggcc tcagtaccct cagtctcctc tccccacatt 1200
gtgtcctgtg ggggtggggt cagccgctgc tctgtacaga accacaggaa ctgatgtgta 1260
tataactatt taatgtggga tatgttcccc tattcctgta tttcccttaa ttcctcctcc 1320
cgacctttтt tacccccccа gttgcagtat ttaactgggc tgggtagggt tgctcagtct 1380
ttgggggagg ttagggactt atcctgtgct tgtaaataaa taaggtcatg actctaaaaa 1440
aaaaaaaagg gcgg 1454

<210> 26
<211> 1121
<212> DNA
<213> Homo sapiens

<220> -
<223> 2328134

<400> 26

gcgggggatg acgccacgga catggtggcc gagaccggcg gggtggggga cgtgtcgcgc 60
ggccgggtgg cctcggtcgg taccctgggc gcggacagct gcctcattag tattcgtacc 120
cacgaggcgg cgcagcgggc cctcggggac agcgagcgtc gcggctatgg cttatcactc 180
gggctacgga gcccacggct ccaagcacag ggcccgggca gccccggatc cccctccgct 240
cttcgatgac acaagcggtg gttattccag ccagcccggg ggatacccag ccacaggagc 300
agacgtggcc ttcagtgtca accacttgct tggggaccca atggccaatg tggctatggc 360
ctatggcagc tccatcgcat cccatgggaa ggacatggtg cacaaggagc tgcaccgttt 420
tgtgtctgtg agcaaactca agtatttttt tgctgtggac acagcctacg tggccaagaa 480
gctagggctg ctggtcttcc cctacacaca ccagaactgg gaagtgcagt acagtcgtga 540
tgctcctctg ccccccggc aagacctcaa cgcccctgac ctctatatcc ccacgatggc 600
cttcattact tacgtgctcc tggctgggat ggcactgggc attcagaaaa ggttctcccc 660
ggaggtgctg ggcctgtgtg caagcacagc gctggtgtgg gtggtgatgg aggtgctggc 720
cctgctcctg ggcctctacc tggccaccgt gcgcagtgac ctgagcacct ttcacctgct 780
ggcctacagt ggctacaaat acgtgggaat gatcctcagt gtgctcacgg ggctgctgtt 840
cggcagcgat ggctactacg tggcgctggc ctggacctca tcggcgctca tgtacttcat 900
tgtgcgctct ttgcggacag cagccctggg ccccgacagc atgggggggcc ccgtcccccg 960
gcagcgtctc cagctctacc tgactctggg agctgcagcc ttccagcccc tcatcatata 1020
ctggctgact ttccacctgg tccggtgacc ccctggcccc agatggcact gagtttttca 1080
ttcattgaag atttgatttc cttgaaaaaa aaaaaaaaag g 1121

<210> 27
<211> 1229
<212> DNA
<213> Homo sapiens

<220> -
<223> 2652271

<400> 27
ctctctgctc cggtgcaggc ccgcaggcgc cctgggctgg gagcaacgcg actgaccgtg 60
gtcgtgggcg gacggcggct gcagcgtgga ggagctgggg tcgctgtggg tcgcgaacag 120
agcccgggac gtgcgcgctt ggtgcacgat cctgaagggg agctccgagg ggcccgggtc 180
tccagggctg ctgcggccat tcccggagcc cggcgcgggg cccgcgagat actggtttag 240
gccgtcccag ggctccgggc gcacccggtg gccgctgctg cagcggaggg agcgcggcgg 300
cgcgggggct cggagacagc gtttctcccg gaagtcttcc tcgggcagca ggtgggaagt 360
gggagccgga gcggcagctg gcagcgttct ctccgcaggt cggcaccatg cgccctgcag 420
ccctgcgcgg ggccctgctg ggctgcctct gcctggcgtt gctttgcctg ggcggtgcgg 480
acaagcgcct gcgtgacaac catgagtgga aaaaactaat tatggttcag cactggcctg 540
agacagtatg cgagaaaatt caaaacgact gtagagaccc tccggattac tggacaatac 600
atggactatg gcccgataaa agtgaaggat gtaatagatc gtggcccttc aatttagaag 660
agattaagga tcttttgcca gaaatgaggg catactggcc tgacgtaatt cactcgtttc 720
ccaatcgcag ccgcttctgg aagcatgagt gggaaaagca tgggacctgc gccgcccagg 780
tggatgcgct caactcccag aagaagtact ttggcagaag cctggaactc tacagggagc 840
tggacctcaa cagtgtgctt ctaaaattgg ggataaaacc atccatcaat tactaccaag 900
ttgcagattt taaagatgcc cttgccagag tatatggagt gatacccaaa atccagtgcc 960
ttccaccaag ccaggatgag gaagtacaga caattggtca gatagaactg tgcctcacta 1020
agcaagacca gcagctgcaa aactgcaccg agccgggga gcagccgtcc cccaagcagg 1080
aagtctggct ggcaaatggg gccgccgaga gccggggtct gagagtctgt gaagatggcc 1140
cagtcttcta tcccccacct aaaaagacca agcattgatg cccaagtttt ggaaatattc 1200
tgttttaaaa agcatgaggt aggcatgtc 1229

<210> 28
<211> 2295
<212> DNA

<213> Homo sapiens

<220> -
<223> 2965248

<400> 28
gtctgcagct ccggccgcca cttgcgcctc tccagcctcc gcaggcccaa ccgccgccag 60
caccatggcc agcaccattt ccgcctacaa ggagaagatg aaggagctgt cggtgctgtc 120
gctcatctgc tcctgcttct acacacagcc gcaccccaat accgtctacc agtacgggga 180
catggaggtg aagcagctgg acaagcgggc ctcaggccag agcttcgagg tcatcctcaa 240
gtccccttct gacctgtccc cagagagccc tatgctctcc tccccaccca agaagaagga 300
cacctccctg gaggagctgc aaaagcggct ggaggcagcc gaggagcgga ggaagacgca 360
ggaggcgcag gtgctgaagc agctggcgga cggcgcgagc acgagcgcga ggtgctgcac 420
aaggcgctgg aggagaataa caacttcagc cgccaggcgg aggagaagct caactacaag 480
atggagctca gcaaggagat ccgcgaggca cacctggccg cactgcgcga gcggctgcgc 540
gagaaggagc tgcacgcggc cgaggtgcgc aggaacaagg agcagcgaga agagatgtcg 600
ggctaagggc ccgggacggg cggcgcccat cctgcgacgg aacacgttcg ggttttggtt 660
ttgtttcgtt cacctctgtc tagatgcaac ttttgttcct cctcccccac cccagccccc 720
agcttcatgc ttctcttccg cactcagccg ccctgccctg tcctcgtggt gagtcgctga 780
ccacggcttc ccctgcagga gccgccgggc gtgagacgcg gtccctcggt gcagacacca 840
ggccgggcgc ggctgggtcc cccgggggcc ctgtgagaga ggtggtggtg accgtggtaa 900
acccagggcg gtggcgtggg atcgcgggtc cttacgctgg gctgtctggt cagcacgtgc 960
aggtcagggc aggtcctctg agccggcgcc cctggccagc aggcgaggct acagtacctg 1020
ctgtctttcc agggggaagg ggctccccat gagggagggg cgacgggggа ggggggtgat 1080
ggtgcctggg agcctgcgtg tgcagccggt gcttgttgaa ctggcaggcg ggtgggtggg 1140
ggctgcagct ttccttaatg tggttgcaca ggggtcctct gagaccacct ggcgtgaggt 1200
ggacaccctg ggccttcctg gaagcctgca gttgggggcc tgccctgagt ctgctgggga 1260
gtgggcattc tctgccaggg acccatgagc aggctgcatg gtctagaggt tgtgggcagc 1320
atggacagtc ccccactcag aagtgcaaga gttccaaaga gcctctggcc caggcccctc 1380
cgtgggacag ccccgccgcc cctcccccacc agggctttgc agatgtcctt gaaagaccca 1440
ccctagagcc ctttggagtg ctggcccctc ctgtgccctc tgccctggtg gaagcggcag 1500
ccacaagtcc tcctcaggga gccccaaggg ggattttgtg ggaccgctgc ccacagatcc 1560
aggtgttgga agggcagcgg gtaaggttcc caagccagcc ccaacaccct tcccacttgg 1620
cacccagagg gggctgtggg tggaggcctg actccaggcc tctcctgccc acaccctctg 1680
ggctgagttc cttctttccc ttggacgccc agtgctggcc ttggaggacg gtcagctgga 1740
ggatggcggt gggggaggct gtctttgtac cactgcagca tcccccactt ctccacggaa 1800
gccccatccc aaagctgctg cctggccccc tgctgtaaag tgtgaagggg gcggctgagt 1860
tctcttagga cccagagcca gggccctcaa cttccatcct gcgggaggcc ttggccgggc 1920
actgccagtg tcttccagag ccacacccag ggaccacggg aggatcctga cccctgcagg 1980
gctcaggggt cagcagggac ccactgcccc atctccctct ccccaccaag acagccccag 2040
aaggagcagc cagctgggat gggaacccaa ggctgtccac atctggcttt tgtgggactc 2100
agaaagggaa gcagaactga gggctgggat attcctcatg gtggcagcgc tcatagcgaa 2160
agcctactgt aatatgcacc catctcatcc acgtagtaaa gtgaacttaa aaattcaatc 2220
aaatgaacaa ttaaataaac acctgtgtgt ttaagacaaa ataaaaatgg aggagaacaa 2280
aaaaaaaggg gcggt 2295

<210> 29
<211> 2215
<212> DNA
<213> Homo sapiens

<220> -
<223> 3057669

<400> 29

cccacgcgtc cgcccacgcg tccgttttca gtagggattt cctgtgacca gacaagttca 60
tctgagagcc agttctcacc actggaattc tcaggaatgg accatgagga catcagtgag 120
tcagtggatg cagcatacaa cctccaggac agttgcctta cagactgtga tgtggaagat 180
gggactatgg atggcaatga tgaggggcac tcctttgaac tttgtccttc tgaagcttct 240
ccttatgtaa ggtcaaggga gagaacctcc tcttcaatag tatttgaaga ttctggctgt 300
gataatgctt ccagtaaaga agagccgaaa actaatcgat tgcatattgg caaccattgt 360
gctaataaac taactgcttt caagcccacc agtagcaaat cttcttctga agctacattg 420
tctatttctc ctccaagacc aaccacttta agtttagatc tcactaaaaa caccacagaa 480
aaactccagc ccagttcacc aaaggtgtat ctttacattc aaatgcagct gtgcagaaaa 540
gaaaacctca aagactggat gaatggacga tgtaccatag aggagagaga gaggagcgtg 600
tgtctgcaca tcttcctgca gatcgcagag gcagtggagt ttcttcacag taaaggactg 660
atgcacaggg acctcaagcc atccaacata ttctttacaa tggatgatgt ggtcaaggtt 720
ggagactttg ggttagtgac tgcaatggac caggatgagg aagagcagac ggttctgacc 780
ccaatgccag cttatgccag acacacagga caagtaggga ccaaactgta tatgagccca 840
gagcagattc atggaaacag ctattctcat aaagtggaca tcttttcttt aggcctgatt 900
ctatttgaat tgctgtatcc attcagcact cagatggaga gagtcaggac cttaactgat 960
gtaagaaatc tcaaatttcc accattattt actcagaaat atccttgtga gtacgtgatg 1020
gttcaagaca tgctctctcc atcccccatg gaacgacctg aagctataaa catcattgaa 1080
aatgctgtat ttgaggactt ggactttcca ggaaaaacag tgctcagaca gaggtctcgc 1140
tccttgagtt catcgggaac aaaacattca agacagtcca acaactccca tagccctttg 1200
ccaagcaatt agccttaagt tgtgctagca accctaatag gtgatgcaga taatagccta 1260
cttcttagaa tatgcctgtc caaaattgca gacttgaaaa gtttgttctt cgctcaattt 1320
ttttgtggac tacttttttt atatcaaatt taagctggat ttgggggcat aacctaattt 1380
gagccaactc ctgagttttg ctatacttaa ggaaagggct atctttgttc tttgttagtc 1440
tcttgaaact ggctgctggc caagctttat agccctcacc atttgcctaa ggaggtagca 1500
gcaatcccta atatatatat atagtgagaa ctaaaatgga tatatttta taatgcagaa 1560
gaaggaaagt ccccctgtgt ggtaactgta ttgttctaga aatatgcttt ctagagatat 1620
gatgattttg aaactgattt ctagaaaaag ctgactccat ttttgtccct ggcgggtaaa 1680
ttaggaatct gcactatttt ggaggacaag tagcacaaac tgtataacgg tttatgtccg 1740
tagttttata gtcctatttg tagcattcaa tagctttatt ccttagatgg ttctagggtg 1800
ggtttacagc tttttgtact tttacctcca ataaagggaa aatgaagctt tttatgtaaa 1860
ttggttgaaa ggtctagttt tgggaggaaa aaagccgtag taagaaatgg atcatatata 1920
ttacaactaa cttcttcaac tatggacttt ttaagcctaa tgaaatctta agtgtcttat 1980
atgtaatcct gtaggttggt acttccccca aactgattat aggtaacagt ttaatcatct 2040
cacttgctaa catgttttta tttttcactg taaatatgtt tatgttttat ttataaaaat 2100
tctgaaatca atccatttgg gttggtggtg tacagaacac acttaagtgt gttaacttgt 2160
gactttcttc aagtctaaat gatttaataa aacttttttt aaattaaaaa aaaaa 2215

<210> 30
<211> 2060
<212> DNA
<213> Homo sapiens

<220> -
<223> 3125156

<400> 30
tccccccctc agcctccccc ccccccactg gcatatggtc ctgccccttc taccagaccc 60
atgggccccc aggcagcccc tcttaccatt cgagggccct cgtctgctgg ccagtccacc 120
cctagtcccc acctggtgcc ttcacctgcc ccatctccag ggcctggtcc ggtacccccc 180
cgccccccag cagcagaacc acccccttgc ctgcgccgag gcgccgcagc tgcagacctg 240
ctctcctcca gcccggagag ccagcatggc ggcactcagt ctcctggggg tgggcagccc 300
ctgctgcagc ccaccaaggt ggatgcagct gagggtcgtc ggccgcaggc cctgcggctg 360
attgagcggg acccctatga gcatcctgag aggctgcggc agttgcagca ggagctggag 420
gcctttcggg gtcagctggg ggatgtggga gctctggaca ctgtctggcg agagctgcaa 480

gatgcgcagg aacatgatgc ccgaggccgt tccatcgcca ttgcccgctg ctactcactg 540
aagaaccggc accaggatgt catgccctat gacagtaacc gtgtggtgct gcgctcaggc 600
aaggatgact acatcaatgc cagctgcgtg gaggggctct ccccatactg ccccccgcta 660
gtggcaaccc aggccccact gcctggcaca gctgctgact tctggctcat ggtccatgag 720
cagaaagtgt cagtcattgt catgctggtt tctgaggctg agatggagaa gcaaaaagtg 780
gcacgctact tccccaccga gaggggccag cccatggtgc acggtgccct gagcctggca 840
ttgagcagcg tccgcagcac cgaaacccat gtggagcgcg tgctgagcct gcagttccga 900
gaccagagcc tcaagcgctc tcttgtgcac ctgcacttcc ccacttggcc tgagttaggc 960
ctgcccgaca gccccagcaa cttgctgcgc ttcatccagg aggtgcacgc acattacctg 1020
catcagcggc cgctgcacac gcccatcatt gtgcactgca gctctggtgt gggccgcacg 1080
ggagcctttg cactgctcta tgcagctgtg caggaggtgg aggctgggaa cggaatccct 1140
gagctgcctc agctggtgcg gcgcatgcgg cagcagagaa agcacatgct gcaggagaag 1200
ctgcacctca ggttctgcta tgaggcagtg gtgagacacg tggagcaggt cctgcagcgc 1260
catggtgtgc ctcctccatg caaacccttg gccagtgcaa gcatcagcca gaagaaccac 1320
cttcctcagg actcccagga cctggtcctc ggtggggatg tgcccatcag ctccatccag 1380
gccaccattg ccaagctcag cattcggcct cctggggggt tggagtcccc ggttgccagc 1440
ttgccaggcc ctgcagagcc cccaggcctc ccgccagcca gcctcccaga gtctacccca 1500
atcccatctt cctcccaaac ccccttttct ccccactacc tgaggctccc cagcctaagg 1560
aggagccgcc agtgcctgaa gcccccagct cggggccccc ctcctcctcc ctggaattgc 1620
tggcctcctt gaccccagag gccttctccc tggacagctc cctgcggggc aaacagcgga 1680
tgagcaagca taactttctg caggcccata acgggcaagg gctgcgggcc acccggccct 1740
ctgacgaccc cctcagcctt ctggatccac tctggacact caacaagacc tgaacaggtt 1800
ttgcctacct ggtccttaca ctacatcatc atcatctcat gcccacctgc ccacacccag 1860
cagagcttct cagtgggcac agtctcttac tcccatttct gctgcctttg gccctgcctg 1920
gcccagcctg caccctgtg gggtggaaat gtactgcagg ctctgggtca ggttctgctc 1980
ctttatggga cccgacattt ttcagctctt tgctattgaa ataataaacc accctgttct 2040
gtgaaaaaaa aaaaaaaaag 2060